SWITCHING THEORY FOR Q-LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Q-learning is a fundamental algorithmic primitive in reinforcement learning. This paper develops a new framework for analyzing constant step-size tabular Q-learning from a switching linear system (SLS) viewpoint. In particular, we derive a stochastic SLS representation of the Q-learning error, and a finite-time error analysis through the joint spectral radius (JSR) of the corresponding SLS model, where the JSR is the exact worst-case exponential rate of the associated SLS. To the best of our knowledge, this is the first convergence rate analysis of standard Q-learning whose leading exponential rate is expressed through the JSR. The resulting rate is tied to the intrinsic worst-case exponential rate of the direct SLS representation and can be sharper than row-sum upper bounds when those bounds are conservative. We further prove that the JSR of Q-learning equals the largest spectral radius among the deterministic-policy modes and give an exact linear programming characterization that can be evaluated to any prescribed accuracy.

1 INTRODUCTION

Q-learning (Watkins and Dayan, 1992) is a foundational algorithm in reinforcement learning (RL) (Sutton and Barto, 1998) for solving discounted Markov decision processes (MDPs) with unknown transition kernels. Its convergence has been studied extensively over the past several decades. Classical analyses primarily establish asymptotic convergence (Tsitsiklis, 1994; Jaakkola et al., 1994; Borkar and Meyn, 2000; Lee and He, 2020). While these results are fundamental, asymptotic convergence alone does not quantify how rapidly the iterates approach the solution, and therefore does not fully characterize the efficiency of the algorithm. This limitation has motivated a growing body of finite-time convergence analyses, which provide explicit bounds on the progress of the iterates toward the optimal Q-function. Recent advances in finite-time analysis include Szepesvári (1998); Kearns and Singh (1999); Even-Dar and Mansour (2003); Beck and Srikant (2012); Wainwright (2019); Qu and Wierman (2020); Li et al. (2022); Chen et al. (2023). Most existing results view Q-learning as a nonlinear stochastic approximation scheme (Kushner and Yin, 2003) and rely on the contraction property of the Bellman optimality operator.

A useful alternative viewpoint is to regard Q-learning as a discrete-time stochastic switching linear system (SLS) (Lin and Antsaklis, 2009; Liberzon, 2003). This perspective was developed by Lee and He (2020); Lee et al. (2023); Lee (2024); Lim and Lee (2024) and was used to prove a final-iterate bound for constant-step-size Q-learning. In that formulation, the error dynamics are affine rather than linear, because the greedy policy selected by the current iterate may differ from an optimal policy. The resulting affine term is controlled through upper and lower comparison systems. More specifically, the upper comparison system remains an SLS, whereas the lower comparison system is linear. Although this comparison-system approach is useful for deriving valid finite-time bounds, it has an inherent limitation: it controls the Q-learning error through auxiliary systems rather than directly exploiting the SLS structure of the original error recursion. Consequently, intrinsic quantities from SLS theory, such as the joint spectral radius (JSR) (Rota and Strang, 1960; Tsitsiklis and Blondel, 1997; Jungers, 2009), are difficult to apply directly to the original Q-learning dynamics.

To overcome this limitation, this paper develops a direct SLS representation of the Q-learning error, where we focus on constant step-size tabular Q-learning. The key observation is that the Bellman maximization error can be represented exactly as an average of action-wise Q-errors under a suitably chosen stochastic policy. Substituting this identity into the Q-learning recursion yields an SLS conditional-mean error system, while the stochastic component remains a martingale difference. The corresponding convergence rate is characterized by the JSR of the resulting SLS. This is the central

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.

distinction from existing Bellman-contraction and comparison system analyses: the rate is attached to the exact direct SLS drift rather than to an auxiliary contraction or a row-sum upper bound (Lee et al., 2023; Lee, 2024). To the best of our knowledge, this is the first convergence rate framework for standard Q-learning whose leading exponential rate is formulated in terms of the JSR of Q-learning. Since the JSR is the exact worst-case exponential rate of the SLS, the resulting transient bound is tied to the intrinsic drift rate of this direct representation and can be sharper than row-sum comparison bounds (Lee et al., 2023; Lee, 2024) when those bounds are conservative. We emphasize that the goal of this paper is to offer a new perspective and a methodological framework for understanding the convergence of Q-learning. The proposed framework is not intended to replace existing analyses based on Bellman contractions or stochastic approximation, nor does it claim uniformly improved finite-time error bounds across all problem instances.

Beyond the finite-time bounds, we establish two structural results for the JSR of Q-learning. Specifically, we show that the JSR equals the largest spectral radius among all the switching modes corresponding to all deterministic policies, and provide an exact linear programming (LP) feasibility characterization of the JSR. The latter can be evaluated to any prescribed accuracy by bisection without enumerating deterministic policies. Together, these results make the direct rate both interpretable and computationally accessible.

2 Preliminaries

2.1 Notation

The set of real numbers is denoted by $\mathbb{R}$; $\mathbb{R}^m$ is the m-dimensional Euclidean space; and $\mathbb{R}^{m\times r}$ is the set of all $m \times r$ real matrices. For a matrix A, $A^\top$ denotes its transpose. For symmetric matrices, $A \succ 0$ and $A \succeq 0$ denote positive definiteness and positive semidefiniteness, and $A \preceq B$ means $B - A \succeq 0$. The identity matrix is denoted by I. For vectors, e_i is the ith standard basis vector, $\mathbf{1}_m$ is the all-ones vector in $\mathbb{R}^m$, and $\otimes$ denotes the Kronecker product. The quantities $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of a symmetric matrix A, respectively. For a finite set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality. We write $\Delta_m := \{q \in \mathbb{R}^m : q_i \geq 0, \sum_{i=1}^m q_i = 1\}$ for the probability simplex in $\mathbb{R}^m$. For a finite matrix family $\mathcal{H} = \{A_1, \dots, A_N\}$, the notation $\mathrm{co}(\mathcal{H})$ denotes the convex hull $\mathrm{co}(\mathcal{H}) := \left\{\sum_{i=1}^N \lambda_i A_i : \lambda_i \geq 0,\ \sum_{i=1}^N \lambda_i = 1\right\}$. For probability distributions p and q on the same finite set $\mathcal{X}$, $\|p - q\|_{\mathrm{TV}} := \frac{1}{2}\sum_{x\in\mathcal{X}} |p(x) - q(x)|$.

For vectors $x, y \in \mathbb{R}^m$, the inequalities $x \geq y$ and $x \gg y$ mean $x_i \geq y_i$ and $x_i > y_i$, respectively, for every coordinate i. In particular, $x \gg 0$ means that every entry of x is strictly positive. A matrix is entrywise positive when each of its entries is strictly positive.

2.2 Switching linear system (SLS)

Let us consider the discrete-time stochastic SLS with a possible affine term (Liberzon, 2003; Lin and Antsaklis, 2009; Shorten et al., 2007)

$$x_{k+1} = A_{\sigma_k} x_k + b_{\sigma_k} + \xi_k,$$

where $\sigma_k \in \{1, 2, \dots, M\}$ is a switching signal, A_{σ_k} is selected from a prescribed family of matrices $\mathcal{H} := \{A_1, A_2, \dots, A_M\}$, which is called a switching family, b_{σ_k} is a mode-dependent affine term, and ξ_k is a disturbance or martingale noise. When $b_{\sigma_k} = 0$, the deterministic part reduces to an SLS,

$$x_{k+1} = A_{\sigma_k} x_k.$$

The worst-case exponential rate of the SLS family is characterized by the joint spectral radius (JSR) (Tsitsiklis and Blondel, 1997; Rota and Strang, 1960; Blondel and Nesterov, 2005; Jungers, 2009), defined as follows.

Definition 1. *For a bounded set of matrices $\mathcal{H} \subset \mathbb{R}^{m\times m}$, its joint spectral radius is*

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

where the value is independent of the chosen submultiplicative norm (Rota and Strang, 1960; Jungers, 2009). When $\mathcal{H}$ is finite, the supremum for each fixed product length is a maximum over products generated by matrices in $\mathcal{H}$. For a finite family $\mathcal{H}$, the notation $\rho(\mathrm{co}(\mathcal{H}))$ means the JSR computed when each factor in a product is allowed to be any convex combination of matrices in $\mathcal{H}$.

Lemma 1. *Consider the SLS*

$$x_{k+1} = A_{\sigma_k} x_k, \qquad A_{\sigma_k} \in \mathcal{H},$$

where $\mathcal{H} \subset \mathbb{R}^{m\times m}$ is finite. If $\rho(\mathcal{H}) < 1$, then $x_k \to 0$ for every initial condition and every switching sequence. Moreover, the convergence is exponential.

The proof is given in Appendix G. For Q-learning (Sutton and Barto, 1998), this perspective is natural because the greedy policy induced by the current iterate plays the role of the switching signal. In a final-iterate analysis, Lee (2024) represented constant-step-size Q-learning under an independent and identically distributed (i.i.d.) observation model by a stochastic affine SLS recursion. The affine term appears because the greedy policy associated with Q_k need not be optimal. We keep the SLS viewpoint, but replace the affine representation with a direct SLS representation obtained from a stochastic policy linearization of the Bellman maximization error.

2.3 Discounted Markov decision processes and Bellman operator

Let us consider a finite discounted Markov decision process (MDP) (Puterman, 1994) with state space $\mathcal{S} = \{1, \dots, |\mathcal{S}|\}$, action space $\mathcal{A} = \{1, \dots, |\mathcal{A}|\}$, transition probability $P(s' \mid s, a)$ for $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, one-step reward $r(s, a, s')$ for $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, expected reward $R(s, a) := \sum_{s'\in\mathcal{S}} P(s' \mid s, a) r(s, a, s'), (s, a) \in \mathcal{S} \times \mathcal{A}$, and discount factor $\gamma \in (0, 1)$. Let Θ denote the set of deterministic stationary policies $\pi : \mathcal{S} \to \mathcal{A}$. A Q-function is viewed as a vector $Q \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$ using a fixed ordering of state-action pairs. We use the compact notation

$$P := \begin{bmatrix} P_1 \\ \vdots \\ P_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times|\mathcal{S}|}, \qquad R := \begin{bmatrix} R(\cdot, 1) \\ \vdots \\ R(\cdot, |\mathcal{A}|) \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|},$$

where $P_a = P(\cdot \mid \cdot, a) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}, a \in \mathcal{A}$. For any stochastic policy $\mu : \mathcal{S} \to \Delta_{|\mathcal{A}|}$, define

$$\Pi^\mu := \begin{bmatrix} \mu(1)^\top \otimes e_1^\top \\ \mu(2)^\top \otimes e_2^\top \\ \vdots \\ \mu(|\mathcal{S}|)^\top \otimes e_{|\mathcal{S}|}^\top \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

Then $P\Pi^\mu \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}$ is the state-action transition matrix under μ. For a deterministic policy $\pi \in \Theta$, we use the same notation Π^π by identifying $\pi(s)$ with its one-hot encoding for each $s \in \mathcal{S}$. For $Q \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$, let us define

$$V_Q(s) := \max_{a\in\mathcal{A}} Q(s, a), \qquad s \in \mathcal{S},$$

and $V_Q := (V_Q(1), \dots, V_Q(|\mathcal{S}|))^\top \in \mathbb{R}^{|\mathcal{S}|}$. The Bellman optimality operator is $F(Q) := R+\gamma P V_Q$. The optimal Q-function Q^* is the unique fixed point of F, i.e., $Q^* = F(Q^*)$, and $V^* := V_{Q^*}$.

Definition 2. *Since $\mathcal{S}$ and $\mathcal{A}$ are finite, let us define*

$$R_{\max} := \max_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}} |r(s, a, s')|.$$

It is well-known that the optimal Q-function then satisfies $\|Q^*\|_\infty \le \frac{R_{\max}}{1-\gamma}$ (Puterman, 1994).

2.4 Q-learning setting

We first consider the standard asynchronous Q-learning recursion (Sutton and Barto, 1998) with a constant step-size α under an i.i.d. observation model. Throughout the paper, we assume that at each time k, a state-action pair (s_k, a_k) is sampled independently with $\mathbb{P}(s_k = s, a_k = a) = d(s, a) := p(s)b(a \mid s), (s, a) \in \mathcal{S} \times \mathcal{A}$, where p is a state sampling distribution and b is a fixed behavior policy. Then, $s'_k \sim P(\cdot \mid s_k, a_k)$ and $r_{k+1} := r(s_k, a_k, s'_k)$ are sampled. The update is

$$Q_{k+1}(s_k, a_k) = Q_k(s_k, a_k) + \alpha \left(r_{k+1} + \gamma \max_{u\in\mathcal{A}} Q_k(s'_k, u) - Q_k(s_k, a_k) \right),$$

while all coordinates $(s,a) \in \mathcal{S} \times \mathcal{A}$ with $(s,a) \neq (s_k, a_k)$ remain unchanged. Let $\{\mathcal{F}_k\}_{k\geq 0}$ be the natural filtration of the Q-learning process

$$\mathcal{F}_0 := \sigma(Q_0), \qquad \mathcal{F}_k := \sigma(Q_0, \{(s_t, a_t, s'_t, r_{t+1}) : 0 \leq t \leq k-1\}), \quad k \geq 1.$$

Then, Q_k is $\mathcal{F}_k$-measurable. Let $D := \mathrm{diag}(d(s,a))_{(s,a)\in\mathcal{S}\times\mathcal{A}} \in \mathbb{R}^{(|\mathcal{S}||\mathcal{A}|)\times(|\mathcal{S}||\mathcal{A}|)}$. For the sampled coordinate, define the random vector

$$\zeta_k := (e_{a_k} \otimes e_{s_k}) \left(r_{k+1} + \gamma \max_{u\in\mathcal{A}} Q_k(s'_k, u) - Q_k(s_k, a_k) \right).$$

Then, we have $\mathbb{E}[\zeta_k \mid \mathcal{F}_k] = D(F(Q_k) - Q_k)$. Accordingly, the martingale-difference noise is defined by

$$w_k := (e_{a_k} \otimes e_{s_k}) \left(r_{k+1} + \gamma \max_{u\in\mathcal{A}} Q_k(s'_k, u) - Q_k(s_k, a_k) \right) - D(F(Q_k) - Q_k). \tag{1}$$

Therefore, the vector form of Q-learning is

$$Q_{k+1} = Q_k + \alpha\{D(F(Q_k) - Q_k) + w_k\},$$

where w_k is defined in Equation (1) and satisfies $\mathbb{E}[w_k \mid \mathcal{F}_k] = 0$ with respect to the natural filtration $\{\mathcal{F}_k\}_{k\geq 0}$ defined above.

Assumption 1. *The following conditions hold throughout the paper.*

(i) $d(s,a) > 0$ *for every* $(s,a) \in \mathcal{S} \times \mathcal{A}$.

(ii) The step-size satisfies $\alpha \in (0,1)$.

(iii) The initial Q-table $Q_0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ *is deterministic.*

For convenience, we define the notations

$$d_{\min} := \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} d(s,a), \qquad d_{\max} := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} d(s,a).$$

Since d is a probability distribution on $|\mathcal{S}||\mathcal{A}|$ coordinates, we have

$$0 < d_{\min} \leq \frac{1}{|\mathcal{S}||\mathcal{A}|}.$$

3 Affine SLS representation of Q-learning

We next recall the affine SLS representation used in earlier switching-system analyses (Lee and He, 2020; Lee et al., 2023; Lee, 2024). For each $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, let π_Q be a deterministic greedy policy satisfying $\pi_Q(s) \in \arg\max_{a\in\mathcal{A}} Q(s,a), s \in \mathcal{S}$. Then

$$V_Q = \Pi^{\pi_Q} Q, \qquad F(Q) = R + \gamma P \Pi^{\pi_Q} Q.$$

Let $\pi_k := \pi_{Q_k}$, and let π^* be an optimal deterministic policy so that $Q^* = R + \gamma P \Pi^{\pi^*} Q^*$. Subtracting the fixed-point equation from the Q-learning recursion gives the stochastic affine SLS recursion

$$(Q_{k+1} - Q^*) = \left(I - \alpha D + \alpha\gamma D P \Pi^{\pi_k}\right)(Q_k - Q^*) + \alpha\gamma D P (\Pi^{\pi_k} - \Pi^{\pi^*}) Q^* + \alpha w_k,$$

where

$$w_k := (e_{a_k} \otimes e_{s_k}) \left(r_{k+1} + \gamma \max_{u\in\mathcal{A}} Q_k(s'_k, u) - Q_k(s_k, a_k) \right) - D(F(Q_k) - Q_k).$$

Equivalently, it can be written as

$$(Q_{k+1} - Q^*) = A_{\pi_k}(Q_k - Q^*) + b_{\pi_k} + \alpha w_k,$$

where $A_\pi := I - \alpha D + \alpha\gamma D P \Pi^\pi$ and $b_\pi := \alpha\gamma D P (\Pi^\pi - \Pi^{\pi^*}) Q^*$. The representation is affine because b_{π_k} is generally nonzero whenever the greedy policy π_k is not optimal. The next section shows that this offset is an artifact of the chosen parametrization: $V_{Q_k} - V^*$ can instead be represented directly as the action-wise error averaged under a suitable stochastic policy.

4 SLS REPRESENTATION OF Q-LEARNING

We now remove the affine offset from the error recursion. The only ingredient needed for this step is the state-wise convexity fact (Goyal and Grand-Clément, 2022), stated and proved in Appendix A.2, Lemma 6: the Bellman maximization error can be written exactly as an average of action-wise Q-errors under a suitable stochastic policy. Substituting this linearization into the Q-learning recursion removes the affine term exactly.

Theorem 1. *Under Assumption 1, let $\{\mathcal{F}_k\}_{k\geq 0}$ be the natural filtration defined in Section 2.4. Then, for each $k \geq 0$, there exists an $\mathcal{F}_k$-measurable stochastic policy μ_k such that*

$$(Q_{k+1} - Q^*) = M_{\mu_k}(Q_k - Q^*) + \alpha w_k,$$

where for any stochastic policy μ,

$$M_\mu := I - \alpha D + \alpha\gamma DP\Pi^\mu,$$

and

$$w_k := (e_{a_k} \otimes e_{s_k}) \left(r_{k+1} + \gamma \max_{u\in\mathcal{A}} Q_k(s'_k, u) - Q_k(s_k, a_k) \right) - D(F(Q_k) - Q_k)$$

and $\mathbb{E}[w_k \mid \mathcal{F}_k] = 0$. Consequently,

$$\mathbb{E}[Q_{k+1} - Q^* \mid \mathcal{F}_k] = M_{\mu_k}(Q_k - Q^*). \tag{2}$$

The proof is given in Appendix H. The corresponding switching family is denoted by

$$\mathcal{M}_\alpha := \{M_\pi := I - \alpha D + \alpha\gamma DP\Pi^\pi : \pi \in \Theta\}.$$

The stochastic-policy matrices in Theorem 1 are connected to the deterministic-policy family by Lemma 7, stated and proved in Appendix A.3, and we can show that it is enough to verify $\rho(\mathcal{M}_\alpha) < 1$ when proving convergence of the SLS; no separate condition on $\rho(\mathrm{co}(\mathcal{M}_\alpha))$ is needed.

For later use, let us define the row-sum rate introduced by Lee et al. (2023); Lee (2024):

$$\rho_{\mathrm{row}} := 1 - \alpha d_{\min}(1 - \gamma).$$

Under Assumption 1, $\rho_{\mathrm{row}} \in (0, 1)$. Its comparison with the switching family of Q-learning is established in Lemma 3, stated and proved in Appendix A.4. We now formalize the deterministic SLS representation associated with Theorem 1:

$$x_{k+1} = M_{\mu_k} x_k, \qquad M_{\mu_k} \in \mathrm{co}(\mathcal{M}_\alpha), \qquad k \geq 0, \tag{3}$$

where the switching sequence may be arbitrary and may select any stochastic-policy matrix generated by Theorem 1. Although this is the deterministic counterpart of the conditional-mean recursion, its stability identifies the drift mechanism that will later be combined with martingale noise estimates for stochastic Q-learning. The following lemma verifies the JSR stability condition in Lemma 1 for the switching family $\mathcal{M}_\alpha$ of Q-learning.

Lemma 4. *Under Assumption 1, the switching family of Q-learning is exponentially stable in the JSR sense:*

$$\rho(\mathcal{M}_\alpha) < 1.$$

In fact,

$$\rho(\mathcal{M}_\alpha) \leq \rho_{\mathrm{row}} = 1 - \alpha d_{\min}(1 - \gamma) < 1.$$

Proof. This follows immediately from Lemma 3 and the inequalities $\alpha \in (0, 1)$, $d_{\min} > 0$, and $\gamma \in (0, 1)$. □

Therefore, by Lemma 4, the deterministic SLS in Equation (3) satisfies the hypothesis of Lemma 1. The corresponding common-norm contraction statement is given as Proposition 1 in Appendix A.5. Consequently, by Proposition 1, the deterministic SLS in Equation (3) converges to zero for every switching sequence. Moreover, by Lemma 3, the row-sum rate is only an upper bound on the switching family of Q-learning $\mathcal{M}_\alpha$ and can be strictly larger than the JSR of $\mathcal{M}_\alpha$.

5 FINITE-TIME ERROR BOUND UNDER I.I.D. OBSERVATION

First, we introduce some constants to quantify the martingale-noise scale and the worst-case growth of products generated by the switching family of Q-learning. They are used throughout the finite-time analysis, and their supporting bounds are established in Appendices A.1 and I. For the stochastic recursion in Section 2.4, let us define

$$W_{\max} := \left(R_{\max} + (1+\gamma) \max \left\{ \|Q_0\|_\infty, \frac{R_{\max}}{1-\gamma} \right\} \right)^2. \tag{4}$$

Appendix A.1, Lemma 5, proves that this quantity bounds the conditional second moment of the martingale noise. For any $\varepsilon > 0$ such that

$$\beta_\varepsilon := \rho(\mathcal{M}_\alpha) + \varepsilon < 1,$$

define the product-growth constant

$$K_{\beta_\varepsilon} := \sup_{\ell \geq 0} \beta_\varepsilon^{-\ell} \sup_{M_0, \ldots, M_{\ell-1} \in \mathrm{co}(\mathcal{M}_\alpha)} \|M_{\ell-1} \cdots M_0\|_\infty, \tag{5}$$

where the product of length zero is the identity.

We now derive a finite-time bound for $\mathbb{E}[\|Q_k - Q^*\|_\infty]$. The complete argument appears in Appendix I. Its main idea is to compare the direct SLS dynamics with a stochastic reference system driven by the same noise, so that the noise cancels pathwise in the residual recursion. For a fixed stochastic policy $\bar{\mu}$, consider

$$x_{k+1} = M_{\bar{\mu}} x_k + \alpha w_k, \qquad x_0 = Q_0 - Q^*.$$

Subtracting this recursion from the direct error recursion gives

$$(Q_{k+1} - Q^*) - x_{k+1} = M_{\mu_k}\big((Q_k - Q^*) - x_k\big) + (M_{\mu_k} - M_{\bar{\mu}}) x_k.$$

Thus, the martingale noise αw_k cancels exactly in the residual. The remaining terms are controlled by the JSR product bound and a covariance estimate for the reference system.

Theorem 2. *Let Assumption 1 hold, and let $W_{\max}$ be defined in Equation (4). Fix any $\varepsilon > 0$ such that*

$$\beta_\varepsilon := \rho(\mathcal{M}_\alpha) + \varepsilon < 1,$$

and let K_{β_ε} be defined in Equation (5). Then, for all $k \geq 0$,

$$\begin{aligned} \mathbb{E}[\|Q_k - Q^*\|_\infty] \leq\ & |\mathcal{S}|\,|\mathcal{A}| \sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}} \left(1 + \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)}\right) \\ & + K_{\beta_\varepsilon} \beta_\varepsilon^k \|Q_0 - Q^*\|_\infty \\ & + 2\alpha\gamma d_{\max} K_{\beta_\varepsilon}^2 k \beta_\varepsilon^{k-1} \|Q_0 - Q^*\|_\infty, \end{aligned} \tag{6}$$

with the convention that $k\beta_\varepsilon^{k-1} = 0$ when $k = 0$.

The proof is given in Appendix I. The first term in Equation (6) is the noise floor, and the remaining terms are transients governed by the JSR rate. For fixed problem data and initialization, the noise floor is of order $\sqrt{\alpha}$ and therefore vanishes as $\alpha \downarrow 0$.

The preceding theorem holds at each fixed step-size. Because K_{β_ε} depends on the gap between β_ε and $\rho(\mathcal{M}_\alpha)$, it may increase as α decreases and thereby enlarge the initial size of the transient terms. Appendix A.6 gives an explicit admissible choice of β_ε for which $K_{\beta_\varepsilon} = 1$, independently of α.

6 ANALYSIS OF THE JSR OF Q-LEARNING

The switching family of Q-learning has additional positive structure that makes its JSR more tractable than the JSR of a generic matrix family. In general, computing or even approximating the JSR is computationally hard, and broad classes of the associated decision and approximation problems are NP-hard (Tsitsiklis and Blondel, 1997; Blondel and Nesterov, 2005). Consequently, numerical methods typically provide upper and lower bounds or approximations rather than an exact value (Jungers, 2009). In contrast, the structure of the switching family of Q-learning permits the exact characterizations below.

Theorem 3. *Under Assumption 1,*

$$\rho(\mathcal{M}_\alpha) = \max_{\pi \in \Theta} \rho(M_\pi), \tag{7}$$

where $\rho(M_\pi)$ denotes the ordinary spectral radius of the single matrix M_π.

The proof is given in Appendix J. Thus, Theorem 3 reduces the JSR of Q-learning to the largest ordinary spectral radius among the deterministic-policy modes, which is a substantial simplification relative to a generic switching family. The next theorem provides an alternative linear programming (LP) characterization that avoids explicit enumeration of deterministic policies.

Theorem 4. *Under Assumption 1, $\rho(\mathcal{M}_\alpha)$ equals the infimum of all β for which there exist $q \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$ and $v \in \mathbb{R}^{|\mathcal{S}|}$ satisfying the following vector inequalities:*

$$q \geq \mathbf{1}_{|\mathcal{S}|\,|\mathcal{A}|}, \tag{8a}$$

$$q \leq \mathbf{1}_{|\mathcal{A}|} \otimes v, \tag{8b}$$

$$(I - \alpha D)q + \alpha\gamma DPv \leq \beta q. \tag{8c}$$

All inequalities are interpreted entrywise. In particular, Equation (8b) is equivalent to $q(s,a) \leq v(s)$ for every $(s,a) \in \mathcal{S} \times \mathcal{A}$.

The proof is given in Appendix K. For fixed β, Equation (8) is an LP feasibility problem with $|\mathcal{S}|\,|\mathcal{A}| + |\mathcal{S}|$ variables. Bisection in β locates the infimum to any prescribed accuracy without enumerating deterministic policies. For small state-action spaces, the LP and bisection procedure can compute the JSR of Q-learning to high numerical accuracy. More importantly, a feasible pair (q, v) serves as a theoretical certificate: it yields a common weighted maximum norm and an explicit finite upper bound for K_{β_ε}. These consequences are proved in Appendix F.

7 Finite-time error bound under Markovian observation

This section applies the ideas from Theorem 2 to Markovian observations and derives a finite-time error bound for single-trajectory Q-learning. Let a behavior policy $b(a \mid s)$ generate

$$s_{k+1} \sim P(\cdot \mid s_k, a_k), \qquad r_{k+1} = r(s_k, a_k, s_{k+1}), \qquad a_{k+1} \sim b(\cdot \mid s_{k+1}),$$

and put $X_k := (s_k, a_k)$. The behavior-induced chain has transition kernel

$$P^b(s', a' \mid s, a) := P(s' \mid s, a) b(a' \mid s').$$

The single-trajectory update is

$$Q_{k+1}(s_k, a_k) = Q_k(s_k, a_k) + \alpha \left(r_{k+1} + \gamma \max_{u \in \mathcal{A}} Q_k(s_{k+1}, u) - Q_k(s_k, a_k) \right), \tag{9}$$

while all other coordinates remain unchanged. The conditional sampling noise is no longer a martingale difference with respect to the stationary mean drift. We control the temporal dependence through the following uniform geometric mixing condition.

Assumption 2. *The behavior-induced chain $X_k = (s_k, a_k)$ is irreducible and aperiodic with stationary distribution d, may start from an arbitrary distribution of X_0, and, for some $t_{\mathrm{mix}} \geq 1$, satisfies*

$$\sup_{x \in \mathcal{S} \times \mathcal{A}} \left\| (P^b)^\ell(x, \cdot) - d \right\|_{\mathrm{TV}} \leq 2^{-\lfloor \ell / t_{\mathrm{mix}} \rfloor}, \qquad \ell \geq 0.$$

For a finite irreducible and aperiodic Markov chain, uniform geometric convergence holds for some finite t_{mix}; the base 2 only fixes a convenient normalization of the mixing-time constant (Levin and Peres, 2017). Conditions of this type are standard in finite-time analyses of Markovian stochastic approximation (Chen et al., 2023).

Under this condition, the direct SLS representation can be combined with a Poisson-equation decomposition of the Markovian noise. The resulting bound retains the same JSR-governed transient terms as in the i.i.d. case, while the temporal dependence contributes a mixing-time term that vanishes with the step-size.

Theorem 5. *Let Assumptions 1 and 2 hold, with d equal to the stationary distribution of the behavior-induced chain. Fix any $\varepsilon > 0$ such that*

$$\beta_\varepsilon := \rho(\mathcal{M}_\alpha) + \varepsilon < 1,$$

and let q and v satisfy Equation (8) with $\beta = \beta_\varepsilon$. Let K_{β_ε} be the product-growth constant from Lemma 9, and let $W_{\max}$ be defined in Equation (4). Then, for every $k \geq 0$,

$$\begin{aligned}\mathbb{E}[\|Q_k - Q^*\|_\infty] \leq & \left(1 + \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)}\right) 128|\mathcal{S}|^2|\mathcal{A}|^2\sqrt{W_{\max}} \\ & \times \left(t_{\mathrm{mix}}\sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)}\right) \\ & + K_{\beta_\varepsilon}\beta_\varepsilon^k\|Q_0 - Q^*\|_\infty \\ & + 2\alpha\gamma d_{\max}K_{\beta_\varepsilon}^2 k\beta_\varepsilon^{k-1}\|Q_0 - Q^*\|_\infty,\end{aligned} \tag{10}$$

where $k\beta_\varepsilon^{k-1} = 0$ when $k = 0$.

The proof is given in Appendix L. It separates the centered fluctuation from the conditional bias through the Poisson equation for P^b. The resulting noise term is $O(t_{\mathrm{mix}}\sqrt{\alpha} + \alpha t_{\mathrm{mix}})$ for fixed problem data and therefore vanishes as $\alpha \downarrow 0$.

8 Numerical Illustration

We summarize the example given in Example 1 of Appendix E. Let us consider $\mathcal{S} = \{1, 2\}$, $\mathcal{A} = \{1, 2\}$, $\gamma = 0.9$, $d = (0.10, 0.40, 0.10, 0.40)$, $P(1 \mid s, a) = P(2 \mid s, a) = \frac{1}{2}$ for every state-action pair. The four deterministic policies generate the switching family of Q-learning, whose exact JSR is

$$\rho(\mathcal{M}_\alpha) = 1 - \frac{55 - 3\sqrt{265}}{400}\alpha < 1 - 0.01\alpha = \rho_{\mathrm{row}}.$$

For $\alpha = 0.30, 0.60, 0.90$, the exact JSR values are 0.995377, 0.990755, and 0.986132, whereas the corresponding row-sum rates are 0.997000, 0.994000, and 0.991000.

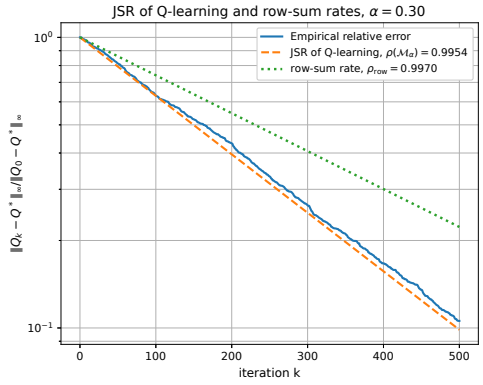
We simulate Q-learning on the same problem. Let $r(s, 1, s') = 1$ and $r(s, 2, s') = -9$ for all $s, s' \in \mathcal{S}$. Action 1 is optimal in both states, with $Q^*(s, 1) = 10$ and $Q^*(s, 2) = 0$, and the initial condition is $Q_0 = 0$. For each step-size, the empirical curve averages 20 independent trajectories and plots the relative error

$$\frac{\|Q_k - Q^*\|_\infty}{\|Q_0 - Q^*\|_\infty}.$$

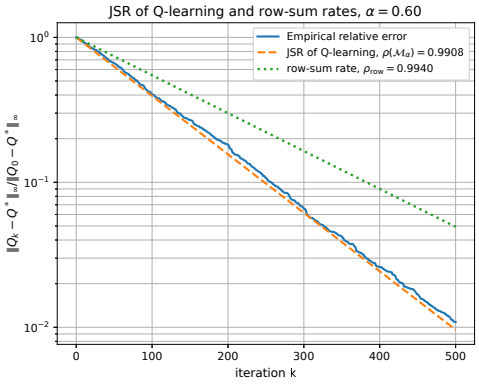
The simulation uses the sampled reward without an additional reward perturbation. Each panel compares the averaged relative error with the normalized curves based on the exact JSR of Q-learning and the row-sum rate. These curves illustrate the separation between the two deterministic rates; they are not empirical sample-complexity bounds.

9 Conclusion

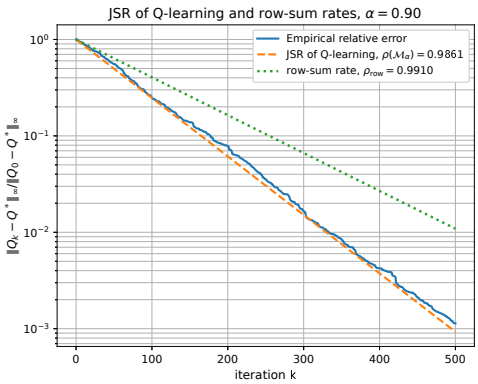
We developed an SLS-based finite-time analysis of Q-learning. The central step is the exact stochastic-policy representation of the Bellman maximization error, which yields an SLS conditional-mean error recursion whose intrinsic deterministic rate is the JSR $\rho(\mathcal{M}_\alpha)$ of the switching family of Q-learning. The standard row-sum rate remains a simple upper bound, but it can be conservative. The resulting bounds use a JSR-induced certificate for the mean drift of Q-learning. The framework also extends to Markovian observations. Beyond the finite-time bounds, the JSR of Q-learning equals the largest spectral radius among the switching modes corresponding to deterministic policies, and its LP characterization can be evaluated by bisection searches. The numerical example shows that this rate can be strictly smaller than the row-sum rate, while also illustrating that a finite-time comparison depends on both the exponential base and the associated prefactor. Overall, the direct-SLS viewpoint complements Bellman-contraction and comparison-system analyses by exposing instance-dependent drift rates and removing comparison-induced transient terms when sharper certificates are available.



(a) $\alpha = 0.30$



(b) $\alpha = 0.60$



(c) $\alpha = 0.90$

Figure 1: Relative Q-learning error for three step-sizes, averaged over 20 independent seeds. Each panel compares the empirical relative error with the curves based on the exact JSR of Q-learning and the row-sum rate, both normalized to start at one.

Reproducibility Statement

This paper primarily makes theoretical contributions. The assumptions and mathematical formulations are stated in the main text, and detailed proofs are provided in the appendix. The examples and simulations are simple illustrations of the theoretical results and can be readily reconstructed from the problem specifications, numerical calculations, and simulation settings provided in Section 8 and Appendix E.

AI Use Statement

Large language model (LLM) tools were used only to aid or polish writing and for retrieval and discovery (e.g., finding related work).

References

Carolyn L Beck and Rayadurgam Srikant. Error bounds for constant step-size Q-learning. *Systems & Control Letters*, 61(12):1203–1208, 2012. doi: 10.1016/j.sysconle.2012.08.014.

Vincent D Blondel and Yurii Nesterov. Computationally efficient approximations of the joint spectral radius. *SIAM Journal on Matrix Analysis and Applications*, 27(1):256–272, 2005. doi: 10.1137/040607009.

Kim C. Border. *Fixed point theorems with applications to economics and game theory*. Cambridge University Press, Cambridge, 1985. doi: 10.1017/CBO9780511625756.

Vivek S Borkar and Sean P Meyn. The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, 2000. doi: 10.1137/S0363012997331639.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. A Lyapunov theory for finite-sample guarantees of Markovian stochastic approximation. *Operations Research*, 72(4):1352–1367, 2023. doi: 10.1287/opre.2022.0249.

Eyal Even-Dar and Yishay Mansour. Learning rates for Q-learning. *Journal of Machine Learning Research*, 5(Dec):1–25, 2003.

Vineet Goyal and Julien Grand-Clément. A first-order approach to accelerated value iteration. *Operations Research*, 71(2):517–535, 2022. doi: 10.1287/opre.2022.2269.

Abhijit Gosavi. Boundedness of iterates in Q-learning. *Systems & Control Letters*, 55(4):347–349, 2006. doi: 10.1016/j.sysconle.2005.08.011.

Jianghai Hu, Jinglai Shen, and Wei Zhang. Generating functions of switched linear systems: analysis, computation, and stability applications. *IEEE Transactions on Automatic Control*, 56(5):1059–1074, 2011. doi: 10.1109/TAC.2010.2067590.

Jianghai Hu, Jinglai Shen, and Donghwan Lee. Resilient stabilization of switched linear control systems against adversarial switching. *IEEE Transactions on Automatic Control*, 62(8):3820–3834, 2017. doi: 10.1109/TAC.2017.2651058.

Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6(6):1185–1201, 1994. doi: 10.1162/neco.1994.6.6.1185.

Raphaël Jungers. *The joint spectral radius: Theory and applications*, volume 385 of *Lecture Notes in Control and Information Sciences*. Springer, Berlin, Heidelberg, 2009. doi: 10.1007/978-3-540-95980-9.

Michael J. Kearns and Satinder P. Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in Neural Information Processing Systems 11*, pages 996–1002, 1999.

Harold Kushner and G. George Yin. *Stochastic approximation and recursive algorithms and applications*, second edition, volume 35 of *Stochastic Modelling and Applied Probability*. Springer, New York, 2003. doi: 10.1007/b97441.

Donghwan Lee. Final iteration convergence bound of Q-learning: Switching system approach. *IEEE Transactions on Automatic Control*, 69(7):4765–4772, 2024. doi: 10.1109/TAC.2024.3355326.

Donghwan Lee and Niao He. A unified switching system perspective and convergence analysis of Q-learning algorithms. In *Advances in Neural Information Processing Systems 33*, pages 15556–15567, 2020.

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. doi: 10.1137/22M1489976.

David A. Levin and Yuval Peres. *Markov chains and mixing times*. Second edition, American Mathematical Society, Providence, RI, 2017. doi: 10.1090/mbk/107.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *IEEE Transactions on Information Theory*, 68(1):448–473, 2022. doi: 10.1109/TIT.2021.3120096.

Daniel Liberzon. *Switching in systems and control*. Birkhäuser, Boston, 2003. doi: 10.1007/978-1-4612-0017-8.

Han-Dong Lim and Donghwan Lee. Finite-time analysis of asynchronous Q-learning under diminishing step-size from control-theoretic view. *IEEE Access*, 12:149916–149939, 2024. doi: 10.1109/ACCESS.2024.3476564.

Hai Lin and Panos J Antsaklis. Stability and stabilizability of switched linear systems: A survey of recent results. *IEEE Transactions on Automatic Control*, 54(2):308–322, 2009. doi: 10.1109/TAC.2008.2012009.

Martin L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, New York, 1994. doi: 10.1002/9780470316887.

Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. In *Proceedings of the Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3185–3205. PMLR, 2020.

Gian-Carlo Rota and Gilbert Strang. A note on the joint spectral radius. *Indagationes Mathematicae*, 22:379–381, 1960. doi: 10.1016/S1385-7258(60)50046-1.

Robert Shorten, Fabian Wirth, Oliver Mason, Kai Wulff, and Christopher King. Stability criteria for switched and hybrid systems. *SIAM Review*, 49(4):545–592, 2007. doi: 10.1137/05063516X.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

Csaba Szepesvári. The asymptotic convergence-rate of Q-learning. In *Advances in Neural Information Processing Systems 10*, pages 1064–1070, 1998.

John N Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning*, 16 (3):185–202, 1994. doi: 10.1023/A:1022689125041.

John N Tsitsiklis and Vincent D Blondel. The lyapunov exponent and joint spectral radius of pairs of matrices are hard—when not impossible—to compute and to approximate. *Mathematics of Control, Signals and Systems*, 10(1):31–40, 1997. doi: 10.1007/BF01219774.

Martin J Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning. Technical report, University of California, Berkeley, arXiv:1905.06265, 2019. doi: 10.48550/arXiv.1905.06265.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3–4):279–292, 1992. doi: 10.1007/BF00992698.

Appendix

A Lemmas and auxiliary results

A.1 Boundedness and noise variance

The first lemma establishes a uniform bound on the iterates and a conditional second-moment bound for the martingale noise used later.

Lemma 5. *Under Assumption 1, let $\{\mathcal{F}_k\}_{k\geq 0}$ be the natural filtration defined in Section 2.4. In particular, Q_0 is deterministic. The Q-learning iterates satisfy*

$$\|Q_k\|_\infty \leq \max\left\{\|Q_0\|_\infty, \frac{R_{\max}}{1-\gamma}\right\}, \qquad \forall k \geq 0.$$

Moreover, for the noise

$$w_k := (e_{a_k} \otimes e_{s_k})\left(r_{k+1} + \gamma \max_{u\in\mathcal{A}} Q_k(s'_k, u) - Q_k(s_k, a_k)\right) - D(F(Q_k) - Q_k),$$

we have

$$\mathbb{E}[\|w_k\|_2^2 \mid \mathcal{F}_k] \leq W_{\max}, \qquad \forall k \geq 0,$$

where $W_{\max}$ is defined in Equation (4).

Proof. The table and noise vectors used below satisfy

$$Q_k, Q_0, \zeta_k, w_k \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}.$$

We prove the two claims separately. Set

$$B_0 := \max\left\{\|Q_0\|_\infty, \frac{R_{\max}}{1-\gamma}\right\}.$$

The absolute boundedness of discounted Q-learning iterates was established by Gosavi (2006, Theorem 1); the sharper invariant bound below follows from the same induction argument. Indeed, if $\|Q_k\|_\infty \leq B_0$, then the sampled target

$$Y_k := r_{k+1} + \gamma \max_{u\in\mathcal{A}} Q_k(s'_k, u)$$

satisfies

$$|Y_k| \leq R_{\max} + \gamma B_0 \leq (1-\gamma)B_0 + \gamma B_0 = B_0.$$

The updated coordinate is

$$Q_{k+1}(s_k, a_k) = (1-\alpha)Q_k(s_k, a_k) + \alpha Y_k,$$

which remains in $[-B_0, B_0]$ because $\alpha \in (0, 1)$, while every other coordinate is unchanged. Since the claim holds at $k = 0$, the same induction proves $\|Q_k\|_\infty \leq B_0$ for every $k \geq 0$.

We now prove the conditional second-moment bound for the noise. Let

$$\zeta_k := (e_{a_k} \otimes e_{s_k})\left(r_{k+1} + \gamma \max_{u\in\mathcal{A}} Q_k(s'_k, u) - Q_k(s_k, a_k)\right).$$

By definition, $w_k = \zeta_k - \mathbb{E}[\zeta_k \mid \mathcal{F}_k]$. For any square-integrable random vector $Z \in \mathbb{R}^m$ and any σ-field $\mathcal{F}$,

$$\mathbb{E}[\|Z - \mathbb{E}[Z \mid \mathcal{F}]\|_2^2 \mid \mathcal{F}] = \mathbb{E}[\|Z\|_2^2 \mid \mathcal{F}] - \|\mathbb{E}[Z \mid \mathcal{F}]\|_2^2 \leq \mathbb{E}[\|Z\|_2^2 \mid \mathcal{F}].$$

Applying this identity with $Z = \zeta_k$ gives

$$\mathbb{E}[\|w_k\|_2^2 \mid \mathcal{F}_k] \leq \mathbb{E}[\|\zeta_k\|_2^2 \mid \mathcal{F}_k].$$

The vector ζ_k has at most one nonzero coordinate. Therefore, its Euclidean norm is exactly the absolute value of the scalar temporal-difference term. By the boundedness just proved, one can derive

$$\begin{aligned}\left|r_{k+1}+\gamma\max_{u\in\mathcal{A}}Q_k(s'_k,u)-Q_k(s_k,a_k)\right| &\le R_{\max}+\gamma\max\left\{\|Q_0\|_\infty,\frac{R_{\max}}{1-\gamma}\right\}\\ &\quad+\max\left\{\|Q_0\|_\infty,\frac{R_{\max}}{1-\gamma}\right\}\\ &= R_{\max}+(1+\gamma)\max\left\{\|Q_0\|_\infty,\frac{R_{\max}}{1-\gamma}\right\}.\end{aligned}$$

Consequently, we have

$$\|\zeta_k\|_2^2\le\left(R_{\max}+(1+\gamma)\max\left\{\|Q_0\|_\infty,\frac{R_{\max}}{1-\gamma}\right\}\right)^2=W_{\max}.$$

Taking conditional expectation on both sides of the above inequality gives the desired bound. □

Note that since Q_0 is deterministic by Assumption 1, $W_{\max}$ is a deterministic finite constant. Since $\max\{\|Q_0\|_\infty, R_{\max}/(1-\gamma)\}\ge R_{\max}/(1-\gamma)$, the crude but useful bound on $W_{\max}$ is $W_{\max}\le 4\max\left\{\|Q_0\|_\infty,\frac{R_{\max}}{1-\gamma}\right\}^2$. In particular, under the common initialization convention $\|Q_0\|_\infty\le R_{\max}/(1-\gamma)$, one has

$$W_{\max}\le\frac{4R_{\max}^2}{(1-\gamma)^2}. \tag{11}$$

We will use the boundedness and noise estimate in Lemma 5 throughout the finite-time analysis.

A.2 Stochastic-policy linearization of the Bellman max

The next lemma converts each state-wise Bellman maximization difference into an exact stochastic-policy average, which supplies the direct SLS representation.

Lemma 6. *For every $Q\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, there exists a stochastic policy $\mu_Q:\mathcal{S}\to\Delta_{|\mathcal{A}|}$ such that*

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

Moreover, μ_Q can be chosen as a measurable function of Q.

Proof. Fix a state $s\in\mathcal{S}$ and write

$$e:=Q-Q^*\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}.$$

We first show that the scalar number

$$y_s:=V_Q(s)-V^*(s)$$

lies between the smallest and largest action-wise errors at state s. Since

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

we have the upper bound

$$\begin{aligned}V_Q(s)&=\max_{a\in\mathcal{A}}\{Q^*(s,a)+e(s,a)\}\\ &\le\max_{a\in\mathcal{A}}Q^*(s,a)+\max_{a\in\mathcal{A}}e(s,a)=V^*(s)+\max_{a\in\mathcal{A}}e(s,a),\end{aligned}$$

which leads to

$$y_s\le\max_{a\in\mathcal{A}}e(s,a).$$

For the lower bound, let $a^*(s)$ be a maximizer of $Q^*(s,\cdot)$. Then

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

Hence, it follows that

$$\min_{a\in\mathcal{A}} e(s,a) \leq y_s \leq \max_{a\in\mathcal{A}} e(s,a).$$

In one dimension, the convex hull of the finite set $\{e(s,a) : a \in \mathcal{A}\}$ is exactly the interval between its minimum and maximum. Therefore y_s can be written as a convex combination of the action-wise errors. We now choose one such convex combination explicitly. Break ties lexicographically and define

$$a_{\min}(s) \in \arg\min_{a\in\mathcal{A}} e(s,a), \qquad a_{\max}(s) \in \arg\max_{a\in\mathcal{A}} e(s,a).$$

If

$$e(s,a_{\min}(s)) = e(s,a_{\max}(s)),$$

then all action-wise errors at state s are equal. In this case y_s must equal that common value, so we may set

$$\mu_Q(a_{\min}(s) \mid s) = 1.$$

If the two values are different, define

$$\lambda_s := \frac{y_s - e(s,a_{\min}(s))}{e(s,a_{\max}(s)) - e(s,a_{\min}(s))}.$$

The bounds above guarantee $0 \leq \lambda_s \leq 1$. Set

$$\mu_Q(a_{\max}(s) \mid s) = \lambda_s, \qquad \mu_Q(a_{\min}(s) \mid s) = 1-\lambda_s,$$

and set all other action probabilities to zero. Then

$$\sum_{a\in\mathcal{A}} \mu_Q(a \mid s) e(s,a) = (1-\lambda_s)e(s,a_{\min}(s)) + \lambda_s e(s,a_{\max}(s)) = y_s.$$

This proves the desired identity at the fixed state s. Repeating the same construction for every state and stacking the resulting scalar identities gives the exact vector–matrix identity, with

$$V_Q - V^* \in \mathbb{R}^{|\mathcal{S}|}, \qquad \Pi^{\mu_Q} \in \mathbb{R}^{|\mathcal{S}|\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

Namely,

$$\begin{aligned} V_Q - V^* &= \begin{bmatrix} V_Q(1) - V^*(1) \\ \vdots \\ V_Q(|\mathcal{S}|) - V^*(|\mathcal{S}|) \end{bmatrix} \\ &= \begin{bmatrix} \sum_{a\in\mathcal{A}} \mu_Q(a \mid 1) e(1,a) \\ \vdots \\ \sum_{a\in\mathcal{A}} \mu_Q(a \mid |\mathcal{S}|) e(|\mathcal{S}|,a) \end{bmatrix} \\ &= \begin{bmatrix} \mu_Q(1)^\top \otimes e_1^\top \\ \mu_Q(2)^\top \otimes e_2^\top \\ \vdots \\ \mu_Q(|\mathcal{S}|)^\top \otimes e_{|\mathcal{S}|}^\top \end{bmatrix} (Q - Q^*) \\ &= \Pi^{\mu_Q}(Q - Q^*). \end{aligned}$$

Finally, the above rule can be made deterministic and unique as a function of Q. Because the action set is finite, lexicographic tie-breaking selects exactly one minimizer and exactly one maximizer at each state. Once these two actions are fixed, the formula above determines the coefficient λ_s uniquely; in the equal-error case, the rule deterministically assigns all mass to the selected action. Thus, for each input Q, the construction returns one well-defined policy rather than a set of possible choices. Since this deterministic rule is implemented by finitely many comparisons of continuous functions of Q together with the explicit formula for λ_s, the resulting policy is measurable as a function of Q. □

A.3 Convex-hull stochastic-policy modes

The policies in Theorem 1 are stochastic, whereas the switching family $\mathcal{M}_\alpha$ is defined through deterministic policies. The following result connects these two descriptions.

Lemma 7. *For every stochastic policy μ, $M_\mu \in \mathrm{co}(\mathcal{M}_\alpha)$. Moreover, when the JSR of $\mathrm{co}(\mathcal{M}_\alpha)$ is interpreted as the JSR over the compact set of all convex combinations of matrices in $\mathcal{M}_\alpha$,*

$$\rho(\mathrm{co}(\mathcal{M}_\alpha)) = \rho(\mathcal{M}_\alpha). \tag{12}$$

Proof. The policy-selection and mode matrices in this proof have dimensions

$$\Pi^\mu, \Pi^\pi \in \mathbb{R}^{|\mathcal{S}|\times(|\mathcal{S}|\,|\mathcal{A}|)}, \qquad M_\mu, M_\pi \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

Fix a stochastic policy μ. A deterministic policy $\pi \in \Theta$ chooses one action at every state. Define

$$c_\pi(\mu) := \prod_{s\in\mathcal{S}} \mu(\pi(s) \mid s), \qquad \pi \in \Theta. \tag{13}$$

Then $c_\pi(\mu) \geq 0$ and $\sum_{\pi\in\Theta} c_\pi(\mu) = 1$. Moreover, for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$, the coefficients in Equation (13) satisfy

$$\sum_{\pi\in\Theta:\,\pi(s)=a} c_\pi(\mu) = \mu(a \mid s). \tag{14}$$

The marginal identity Equation (14) implies

$$\Pi^\mu = \sum_{\pi\in\Theta} c_\pi(\mu)\Pi^\pi. \tag{15}$$

By the linear dependence of $M_\mu = I - \alpha D + \alpha\gamma DP\Pi^\mu$ on Π^μ, Equation (15) gives

$$M_\mu = \sum_{\pi\in\Theta} c_\pi(\mu) M_\pi. \tag{16}$$

Thus $M_\mu \in \mathrm{co}(\mathcal{M}_\alpha)$.

It remains to compare the JSRs. We use the standard convex-hull invariance property of the JSR: for any bounded family of matrices, taking the convex hull does not change its joint spectral radius (Jungers, 2009, Section 1.2.2.7); see also Rota and Strang (1960). Applying this property to the finite family $\mathcal{M}_\alpha$ gives Equation (12), which completes the proof. □

A.4 Row-sum upper bound for the JSR of Q-learning

We compare the JSR of Q-learning with the row-sum rate

$$\rho_{\mathrm{row}} = 1 - \alpha d_{\min}(1-\gamma),$$

which is used in the main text. Under Assumption 1, this rate belongs to $(0, 1)$. The following lemma relates it to every stochastic-policy mode and, consequently, to the switching family of Q-learning.

Lemma 3. *Under Assumption 1, the row-sum rate is an upper bound on the JSR of Q-learning:*

$$\rho(\mathcal{M}_\alpha) \leq \rho_{\mathrm{row}}.$$

More generally, for every stochastic policy μ,

$$\|M_\mu\|_\infty \leq \rho_{\mathrm{row}}.$$

Proof. The matrices used below satisfy

$$D, M_\mu \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}, \quad P \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times|\mathcal{S}|}, \quad \Pi^\mu \in \mathbb{R}^{|\mathcal{S}|\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

Note that the matrices D, P, and Π^μ are nonnegative. Moreover, $1-\alpha d(s,a) \geq 0$ because $0 < \alpha < 1$ and $d(s,a) \leq 1$. Hence every entry of M_μ is nonnegative. For the row of $P\Pi^\mu$ indexed by (s,a),

$$\begin{aligned}\sum_{(\bar{s},\bar{a})\in\mathcal{S}\times\mathcal{A}} [P\Pi^\mu]_{(s,a),(\bar{s},\bar{a})} &= \sum_{\bar{s}\in\mathcal{S}} P(\bar{s} \mid s,a) \sum_{\bar{a}\in\mathcal{A}} \mu(\bar{a} \mid \bar{s}) \\ &= \sum_{\bar{s}\in\mathcal{S}} P(\bar{s} \mid s,a) \\ &= 1.\end{aligned}$$

Thus the corresponding row of M_μ sums to

$$\begin{aligned}&\sum_{(\bar{s},\bar{a})\in\mathcal{S}\times\mathcal{A}} [M_\mu]_{(s,a),(\bar{s},\bar{a})} \\ &\quad= 1 - \alpha d(s,a) + \alpha\gamma d(s,a) \sum_{(\bar{s},\bar{a})\in\mathcal{S}\times\mathcal{A}} [P\Pi^\mu]_{(s,a),(\bar{s},\bar{a})} \\ &\quad= 1 - \alpha d(s,a) + \alpha\gamma d(s,a) \\ &\quad= 1 - \alpha d(s,a)(1-\gamma) \\ &\quad\leq 1 - \alpha d_{\min}(1-\gamma) \\ &\quad= \rho_{\text{row}}.\end{aligned}$$

For a nonnegative matrix, the induced infinity norm is the maximum row sum. Consequently,

$$\begin{aligned}\|M_\mu\|_\infty &= \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \sum_{(\bar{s},\bar{a})\in\mathcal{S}\times\mathcal{A}} [M_\mu]_{(s,a),(\bar{s},\bar{a})} \\ &\leq \rho_{\text{row}}\end{aligned}$$

for every stochastic policy μ, and hence for every deterministic policy. By the JSR definition in Definition 1 and submultiplicativity of the induced infinity norm,

$$\begin{aligned}\rho(\mathcal{M}_\alpha) &= \lim_{\ell\to\infty} \max_{M_0,\ldots,M_{\ell-1}\in\mathcal{M}_\alpha} \|M_{\ell-1}\cdots M_0\|_\infty^{1/\ell} \\ &\leq \limsup_{\ell\to\infty} \max_{M_0,\ldots,M_{\ell-1}\in\mathcal{M}_\alpha} \left(\prod_{i=0}^{\ell-1} \|M_i\|_\infty\right)^{1/\ell} \\ &\leq \limsup_{\ell\to\infty} \left(\rho_{\text{row}}^\ell\right)^{1/\ell} \\ &= \rho_{\text{row}}.\end{aligned}$$

This proves both claims. □

The next lemma turns the one-step row-sum estimate into a uniform bound for all products, including convex-combination modes.

Lemma 4. *For every $\alpha \in (0,1)$, the corresponding switching family of Q-learning satisfies*

$$\rho(\mathcal{M}_\alpha) \leq 1 - \alpha d_{\min}(1-\gamma) < 1,$$

and, for every $\ell \geq 0$,

$$\sup_{M_0,\ldots,M_{\ell-1}\in\text{co}(\mathcal{M}_\alpha)} \|M_{\ell-1}\cdots M_0\|_\infty \leq (1 - \alpha d_{\min}(1-\gamma))^\ell. \tag{17}$$

Proof. Every mode appearing in the products below satisfies

$$M_i, M_\pi \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

The JSR inequality follows from Lemma 3. For the product estimate, Lemma 7 implies that, for each $M_i \in \mathrm{co}(\mathcal{M}_\alpha)$, there are coefficients $c_{i,\pi} \geq 0$, $\pi \in \Theta$, such that

$$M_i = \sum_{\pi\in\Theta} c_{i,\pi} M_\pi, \qquad \sum_{\pi\in\Theta} c_{i,\pi} = 1.$$

Convexity of the induced infinity norm and Lemma 3 therefore give

$$\begin{aligned}
\|M_i\|_\infty &= \left\| \sum_{\pi\in\Theta} c_{i,\pi} M_\pi \right\|_\infty \\
&\leq \sum_{\pi\in\Theta} c_{i,\pi} \|M_\pi\|_\infty \\
&\leq \sum_{\pi\in\Theta} c_{i,\pi} \left(1 - \alpha d_{\min}(1-\gamma)\right) \\
&= 1 - \alpha d_{\min}(1-\gamma).
\end{aligned}$$

For $\ell \geq 1$, submultiplicativity then yields

$$\begin{aligned}
\|M_{\ell-1}\cdots M_0\|_\infty &\leq \|M_{\ell-1}\|_\infty \|M_{\ell-2}\cdots M_0\|_\infty \\
&\leq \prod_{i=0}^{\ell-1} \|M_i\|_\infty \\
&\leq \prod_{i=0}^{\ell-1} \left(1 - \alpha d_{\min}(1-\gamma)\right) \\
&= \left(1 - \alpha d_{\min}(1-\gamma)\right)^\ell.
\end{aligned}$$

For $\ell = 0$, both sides equal one. Taking the supremum over all admissible products proves Equation (17). □

A.5 Direct common-norm rate

The following proposition converts the JSR gap into a common norm that contracts every stochastic-policy mode.

Proposition 1. *Consider the discrete SLS in Equation (3). For every $\varepsilon > 0$ such that*

$$\beta_\varepsilon := \rho(\mathcal{M}_\alpha) + \varepsilon < 1, \tag{18}$$

there exists a norm p_ε on $\mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$ such that

$$p_\varepsilon(M_\mu x) \leq \beta_\varepsilon p_\varepsilon(x), \qquad \forall x \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \tag{19}$$

for every stochastic policy μ. Hence p_ε is a common Lyapunov norm for Equation (3); every trajectory satisfies

$$p_\varepsilon(x_k) \leq \beta_\varepsilon^k p_\varepsilon(x_0), \qquad k \geq 0, \tag{20}$$

and therefore $x_k \to 0$ exponentially for every switching sequence.

Proof. The vectors and mode matrices below satisfy

$$x \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \qquad M_\pi, M_\mu \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

By assumption, $\beta_\varepsilon < 1$. A standard extremal-norm consequence of the JSR theory says that, for every number strictly larger than the JSR of a finite matrix family, there exists a norm that contracts

all matrices in the family by that number (Rota and Strang, 1960; Jungers, 2009). A concrete example is the JSR-based norm constructed in Lemma 3. Applying the extremal-norm result to the finite family $\mathcal{M}_\alpha$ gives a norm p_ε satisfying

$$p_\varepsilon(M_\pi x) \le \beta_\varepsilon p_\varepsilon(x), \qquad \forall x \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \forall \pi \in \Theta. \tag{21}$$

Now take an arbitrary stochastic policy μ. By Lemma 7,

$$M_\mu = \sum_{\pi\in\Theta} c_\pi(\mu) M_\pi, \qquad c_\pi(\mu) \ge 0, \qquad \sum_{\pi\in\Theta} c_\pi(\mu) = 1. \tag{22}$$

Using Equation (22), the triangle inequality, positive homogeneity of the norm, and then Equation (21), we get

$$\begin{aligned} p_\varepsilon(M_\mu x) &= p_\varepsilon\left(\sum_{\pi\in\Theta} c_\pi(\mu) M_\pi x\right) \\ &\le \sum_{\pi\in\Theta} c_\pi(\mu) p_\varepsilon(M_\pi x) \\ &\le \sum_{\pi\in\Theta} c_\pi(\mu)\beta_\varepsilon p_\varepsilon(x) = \beta_\varepsilon p_\varepsilon(x), \end{aligned} \tag{23}$$

which is Equation (19). Now let $x_{k+1} = M_{\mu_k} x_k$ be any trajectory of Equation (3). Repeatedly applying Equation (23) gives Equation (20). Because $\beta_\varepsilon < 1$ and p_ε is a norm on a finite-dimensional space, Equation (20) implies $x_k \to 0$ exponentially. This completes the proof. □

A.6 Product-growth consequences

For any $\varepsilon > 0$ such that

$$\beta_\varepsilon := \rho(\mathcal{M}_\alpha) + \varepsilon < 1,$$

define the product-growth constant

$$K_{\beta_\varepsilon} := \sup_{\ell\ge 0} \beta_\varepsilon^{-\ell} \sup_{M_0,\dots,M_{\ell-1}\in \mathrm{co}(\mathcal{M}_\alpha)} \|M_{\ell-1}\cdots M_0\|_\infty ,$$

where the product of length zero is the identity. We first show that any rate, β_ε, strictly above the JSR has a finite product-growth constant $K_{\beta_\varepsilon} = 1$ whenever the rate dominates the row-sum bound.

Lemma 1. *Let Assumption 1 hold, fix $\alpha \in (0,1)$, and choose any $\varepsilon > 0$ such that*

$$\beta_\varepsilon := \rho(\mathcal{M}_\alpha) + \varepsilon < 1.$$

Then $K_{\beta_\varepsilon} < \infty$ for this fixed pair (α, ε). If, in addition,

$$\beta_\varepsilon \ge 1 - \alpha d_{\min}(1-\gamma),$$

then $K_{\beta_\varepsilon} = 1$. In particular, under this additional condition, K_{β_ε} is independent of α and ε.

Proof. By Lemma 7,

$$\rho(\mathrm{co}(\mathcal{M}_\alpha)) = \rho(\mathcal{M}_\alpha) < \beta_\varepsilon.$$

Choose $\eta \in (\rho(\mathrm{co}(\mathcal{M}_\alpha)), \beta_\varepsilon)$. For every product length $\ell \ge 0$, define

$$a_\ell := \sup_{M_0,\dots,M_{\ell-1}\in \mathrm{co}(\mathcal{M}_\alpha)} \|M_{\ell-1}\cdots M_0\|_\infty, \qquad a_0 := 1,$$

where

$$M_0, \dots, M_{\ell-1} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

The sequence is submultiplicative:

$$a_{\ell+j} \le a_\ell a_j, \qquad \ell, j \ge 0.$$

By the JSR definition in Definition 1,

$$\lim_{\ell\to\infty} a_\ell^{1/\ell} = \rho(\mathrm{co}(\mathcal{M}_\alpha)) < \eta.$$

Hence there exists an integer $L_\eta \geq 1$ such that

$$a_\ell^{1/\ell} \leq \eta, \qquad \ell \geq L_\eta,$$

or equivalently, $a_\ell \leq \eta^\ell$ for every $\ell \geq L_\eta$. Define the finite-length correction

$$C_\eta := \max\left\{1, \max_{0\leq j<L_\eta} \eta^{-j} a_j\right\} < \infty.$$

The definition of C_η controls the finitely many lengths below L_η, while the preceding tail estimate controls all larger lengths. Therefore,

$$a_\ell \leq C_\eta \eta^\ell, \qquad \ell \geq 0.$$

Consequently, Equation (5) gives the detailed bound

$$\begin{aligned} K_{\beta_\varepsilon} &= \sup_{\ell\geq 0} \beta_\varepsilon^{-\ell} a_\ell \\ &\leq C_\eta \sup_{\ell\geq 0} \left(\frac{\eta}{\beta_\varepsilon}\right)^\ell \\ &= C_\eta < \infty, \end{aligned}$$

because $0 < \eta/\beta_\varepsilon < 1$, so the supremum of the geometric sequence over $\ell \in \{0, 1, 2, \ldots\}$ is attained at $\ell = 0$ and equals one. This conclusion is pointwise in (α, ε), because the constants C_η and η may depend on the fixed pair. Under the additional inequality, Lemma 3 and convexity of the induced infinity norm imply

$$\|M_i\|_\infty \leq 1 - \alpha d_{\min}(1-\gamma) \leq \beta_\varepsilon, \qquad M_i \in \mathrm{co}(\mathcal{M}_\alpha).$$

Therefore, for every $\ell \geq 0$,

$$\beta_\varepsilon^{-\ell} \|M_{\ell-1} \cdots M_0\|_\infty \leq \left(\frac{1 - \alpha d_{\min}(1-\gamma)}{\beta_\varepsilon}\right)^\ell \leq 1.$$

Taking the supremum in Equation (5) gives $K_{\beta_\varepsilon} \leq 1$. The product of length zero gives the reverse inequality, so $K_{\beta_\varepsilon} = 1$, independently of α and ε. □

The following proposition chooses β_ε explicitly from $d_{\min}$, γ, and α, without requiring the exact value of $\rho(\mathcal{M}_\alpha)$.

Proposition 2. *Let Assumption 1 hold and choose*

$$\beta_\varepsilon := 1 - \frac{d_{\min}(1-\gamma)}{2}\alpha.$$

Then

$$\beta_\varepsilon > \rho(\mathcal{M}_\alpha), \qquad K_{\beta_\varepsilon} = 1,$$

where K_{β_ε} is defined in Equation (5).

Proof. By Lemma 3,

$$\rho(\mathcal{M}_\alpha) \leq 1 - \alpha d_{\min}(1-\gamma) < 1 - \frac{d_{\min}(1-\gamma)}{2}\alpha = \beta_\varepsilon < 1.$$

Thus this choice is admissible and satisfies the additional condition in Lemma 1, which gives $K_{\beta_\varepsilon} = 1$. □

Using this explicit rate in the reference-system estimate yields the following step-size-dependent error bound.

Proposition 3. *Let Assumption 1 hold and consider the i.i.d. Q-learning recursion in Section 2.4. Let $W_{\max}$ be defined in Equation (4). Choose*

$$\beta_\varepsilon := 1 - \frac{d_{\min}(1-\gamma)}{2}\alpha.$$

Then, for every $k \geq 0$,

$$\begin{aligned}\mathbb{E}[\|Q_k - Q^*\|_\infty] \leq |\mathcal{S}|\,|\mathcal{A}|\sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}}\left(1 + \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)}\right) \\ + \left(1 - \frac{d_{\min}(1-\gamma)}{2}\alpha\right)^k \|Q_0 - Q^*\|_\infty \\ + 2\alpha\gamma d_{\max} k \left(1 - \frac{d_{\min}(1-\gamma)}{2}\alpha\right)^{k-1} \|Q_0 - Q^*\|_\infty,\end{aligned} \tag{24}$$

where $k(1 - d_{\min}(1-\gamma)\alpha/2)^{k-1} = 0$ when $k = 0$. If the MDP, d, and Q_0 are fixed independently of α, the first term on the right-hand side is $O(\sqrt{\alpha})$ as $\alpha \downarrow 0$.

Proof. By Proposition 2, the selected β_ε exceeds $\rho(\mathcal{M}_\alpha)$ and satisfies $K_{\beta_\varepsilon} = 1$. Substituting these values into Equation (6) proves Equation (24). □

B JSR-BASED LYAPUNOV FUNCTION CONSTRUCTION

The next lemma constructs a product-based Lyapunov norm and establishes the inequalities, that can be used in the finite-time analysis.

Lemma 2. *Let*

$$\mathcal{H} := \{A_1, A_2, \ldots, A_M\} \subset \mathbb{R}^{m\times m}.$$

Fix any $\varepsilon > 0$ such that

$$\beta_\varepsilon := \rho(\mathcal{H}) + \varepsilon \in (0, 1).$$

For a sequence of switching modes $\sigma = (\sigma_1, \ldots, \sigma_k) \in \{1, \ldots, M\}^k$, write

$$A_\sigma := A_{\sigma_k} \cdots A_{\sigma_1},$$

and for $k = 0$ interpret $\{1, \ldots, M\}^0$ as the singleton empty word and $A_\sigma = I$. For each integer $t \geq 0$, define the piecewise quadratic function

$$V_\varepsilon^t(x) := \sum_{k=0}^{t} \beta_\varepsilon^{-2k} \max_{\sigma\in\{1,2,\ldots,M\}^k} \|A_\sigma x\|_2^2, \qquad x \in \mathbb{R}^m.$$

Then the following statements hold.

1. *For every $t \geq 0$,*

$$V_\varepsilon^{t+1}(x) \geq \|x\|_2^2 + \beta_\varepsilon^{-2} \max_{i\in\{1,\ldots,M\}} V_\varepsilon^t(A_i x), \qquad \forall x \in \mathbb{R}^m. \tag{25}$$

2. *For every $t \geq 0$, every $x \in \mathbb{R}^m$, and every $\lambda \in \mathbb{R}$,*

$$V_\varepsilon^t(\lambda x) = |\lambda|^2 V_\varepsilon^t(x),$$

and

$$V_\varepsilon^t(x) \leq V_\varepsilon^{t+1}(x).$$

3. There exists a constant $C_\varepsilon > 0$ such that

$$\|x\|_2^2 \le V_\varepsilon^t(x) \le C_\varepsilon \|x\|_2^2, \qquad \forall x \in \mathbb{R}^m, \quad \forall t \ge 0. \tag{26}$$

4. For every $x \in \mathbb{R}^m$, the limit

$$V_\varepsilon^\infty(x) := \lim_{t\to\infty} V_\varepsilon^t(x)$$

exists and is finite. Moreover,

$$\|x\|_2^2 \le V_\varepsilon^\infty(x) \le C_\varepsilon \|x\|_2^2, \qquad \forall x \in \mathbb{R}^m. \tag{27}$$

5. The function

$$p_\varepsilon(x) := \sqrt{V_\varepsilon^\infty(x)}$$

is a norm on $\mathbb{R}^m$.

6. The function V_ε^∞ satisfies the stronger Lyapunov inequality

$$V_\varepsilon^\infty(A_i x) \le \beta_\varepsilon^2 \left(V_\varepsilon^\infty(x) - \|x\|_2^2\right) \le \beta_\varepsilon^2 V_\varepsilon^\infty(x), \qquad \forall x \in \mathbb{R}^m, \quad \forall i \in \{1,\dots,M\}.$$

Equivalently,

$$p_\varepsilon(A_i x) \le \beta_\varepsilon p_\varepsilon(x), \qquad \forall x \in \mathbb{R}^m, \quad \forall i \in \{1,\dots,M\}.$$

7. Consequently,

$$\|A_i\|_{p_\varepsilon} \le \beta_\varepsilon, \qquad \forall i \in \{1,\dots,M\},$$

where $\|A_i\|_{p_\varepsilon}$ is the induced matrix norm generated by p_ε,

$$\|A_i\|_{p_\varepsilon} := \sup_{x\in\mathbb{R}^m,\ x\neq 0} \frac{p_\varepsilon(A_i x)}{p_\varepsilon(x)}.$$

Therefore, the induced-norm bound yields

$$\rho(\mathcal{H}) \le \beta_\varepsilon.$$

Proof. Throughout this proof,

$$x \in \mathbb{R}^m, \qquad A_i, A_\sigma \in \mathbb{R}^{m\times m}.$$

We prove the statements one by one.

Proof of (1). By definition,

$$V_\varepsilon^{t+1}(x) = \sum_{k=0}^{t+1} \beta_\varepsilon^{-2k} \max_{\sigma\in\{1,\dots,M\}^k} \|A_\sigma x\|_2^2.$$

The $k = 0$ term is $\|x\|_2^2$. For $k \ge 1$, write $k = j + 1$. A word of length $j + 1$ may be described by its first applied index $i \in \{1,\dots,M\}$ and the remaining word $\tau \in \{1,\dots,M\}^j$. Therefore, decomposing by the first applied index gives

$$\begin{aligned} V_\varepsilon^{t+1}(x) &= \|x\|_2^2 + \sum_{j=0}^{t} \beta_\varepsilon^{-2(j+1)} \max_{i\in\{1,\dots,M\}} \max_{\tau\in\{1,\dots,M\}^j} \|A_\tau A_i x\|_2^2 \\ &= \|x\|_2^2 + \beta_\varepsilon^{-2} \sum_{j=0}^{t} \beta_\varepsilon^{-2j} \max_{i\in\{1,\dots,M\}} \max_{\tau\in\{1,\dots,M\}^j} \|A_\tau A_i x\|_2^2. \end{aligned}$$

Since the summands are nonnegative, the sum of the maxima is at least the maximum of the sums. Hence

$$V_\varepsilon^{t+1}(x) \geq \|x\|_2^2 + \beta_\varepsilon^{-2} \max_{i\in\{1,\dots,M\}} \sum_{j=0}^{t} \beta_\varepsilon^{-2j} \max_{\tau\in\{1,\dots,M\}^j} \|A_\tau A_i x\|_2^2$$
$$= \|x\|_2^2 + \beta_\varepsilon^{-2} \max_{i\in\{1,\dots,M\}} V_\varepsilon^t(A_i x).$$

This proves Equation (25).

Proof of (2). Absolute homogeneity follows from the homogeneity of the Euclidean norm:

$$V_\varepsilon^t(\lambda x) = \sum_{k=0}^{t} \beta_\varepsilon^{-2k} \max_{\sigma\in\{1,\dots,M\}^k} \|A_\sigma(\lambda x)\|_2^2 = |\lambda|^2 V_\varepsilon^t(x).$$

The monotonicity $V_\varepsilon^t(x) \leq V_\varepsilon^{t+1}(x)$ holds because V_ε^{t+1} contains all terms of V_ε^t plus one additional nonnegative term.

Proof of (3). The lower bound is immediate from the $k = 0$ term:

$$V_\varepsilon^t(x) \geq \|x\|_2^2.$$

For the upper bound, choose η such that

$$\rho(\mathcal{H}) < \eta < \beta_\varepsilon.$$

By the definition of the JSR, there exists an integer $K \geq 1$ such that

$$\max_{\sigma\in\{1,\dots,M\}^k} \|A_\sigma\|_2^{1/k} \leq \eta, \qquad \forall k \geq K.$$

Thus, for all $k \geq K$,

$$\max_{\sigma\in\{1,\dots,M\}^k} \|A_\sigma\|_2 \leq \eta^k.$$

Now define

$$C_0 := \max\left\{1, \max_{0\leq k\leq K-1} \eta^{-k} \max_{\sigma\in\{1,\dots,M\}^k} \|A_\sigma\|_2\right\}.$$

Then, for every $k \geq 0$,

$$\max_{\sigma\in\{1,\dots,M\}^k} \|A_\sigma\|_2 \leq C_0 \eta^k.$$

Therefore, the product bound gives the estimate

$$V_\varepsilon^t(x) = \sum_{k=0}^{t} \beta_\varepsilon^{-2k} \max_{\sigma\in\{1,\dots,M\}^k} \|A_\sigma x\|_2^2$$
$$\leq \sum_{k=0}^{t} \beta_\varepsilon^{-2k} \left(\max_{\sigma\in\{1,\dots,M\}^k} \|A_\sigma\|_2\right)^2 \|x\|_2^2$$
$$\leq C_0^2 \sum_{k=0}^{t} \left(\frac{\eta}{\beta_\varepsilon}\right)^{2k} \|x\|_2^2$$
$$\leq C_0^2 \sum_{k=0}^{\infty} \left(\frac{\eta}{\beta_\varepsilon}\right)^{2k} \|x\|_2^2.$$

Since $\eta/\beta_\varepsilon < 1$, the geometric series converges. Setting

$$C_\varepsilon := \frac{C_0^2}{1 - (\eta/\beta_\varepsilon)^2}$$

gives Equation (26).

Proof of (4). By (2), the sequence $V_\varepsilon^t(x)$ is nondecreasing in t. By (3), it is bounded above by $C_\varepsilon \|x\|_2^2$. Hence

$$V_\varepsilon^\infty(x) := \lim_{t\to\infty} V_\varepsilon^t(x)$$

exists and is finite. Passing to the limit in the bounds of (3) proves Equation (27).

Proof of (5). For each $k \geq 1$, define

$$\nu_k(x) := \beta_\varepsilon^{-k} \max_{\sigma\in\{1,\dots,M\}^k} \|A_\sigma x\|_2,$$

and define $\nu_0(x) := \|x\|_2$. Each ν_k is a seminorm because it is the pointwise maximum of seminorms, and ν_0 is a norm. For finite t,

$$V_\varepsilon^t(x) = \sum_{k=0}^{t} \nu_k(x)^2, \qquad p_\varepsilon^t(x) := \sqrt{V_\varepsilon^t(x)} = \left(\sum_{k=0}^{t} \nu_k(x)^2\right)^{1/2}.$$

The function p_ε^t is a norm. Positivity follows from the ν_0 term, absolute homogeneity follows from the homogeneity of each ν_k, and the triangle inequality follows from Minkowski's inequality applied to the vector

$$(\nu_0(x), \nu_1(x), \dots, \nu_t(x)) \in \mathbb{R}^{t+1}.$$

Now

$$p_\varepsilon(x) = \sqrt{V_\varepsilon^\infty(x)} = \lim_{t\to\infty} p_\varepsilon^t(x),$$

and the limit is finite by (4). Since

$$p_\varepsilon^t(x+y) \leq p_\varepsilon^t(x) + p_\varepsilon^t(y)$$

for every t, letting $t \to \infty$ gives

$$p_\varepsilon(x+y) \leq p_\varepsilon(x) + p_\varepsilon(y).$$

Absolute homogeneity is inherited from (2), and positive definiteness follows from

$$p_\varepsilon(x)^2 = V_\varepsilon^\infty(x) \geq \|x\|_2^2.$$

Thus p_ε is a norm.

Proof of (6). Using (1),

$$V_\varepsilon^{t+1}(x) \geq \|x\|_2^2 + \beta_\varepsilon^{-2} \max_{i\in\{1,\dots,M\}} V_\varepsilon^t(A_i x).$$

Therefore, rearranging the finite-horizon inequality gives

$$\max_{i\in\{1,\dots,M\}} V_\varepsilon^t(A_i x) \leq \beta_\varepsilon^2 \left(V_\varepsilon^{t+1}(x) - \|x\|_2^2\right).$$

Fixing i, we have

$$V_\varepsilon^t(A_i x) \leq \beta_\varepsilon^2 \left(V_\varepsilon^{t+1}(x) - \|x\|_2^2\right).$$

Letting $t \to \infty$ and using monotone convergence of both sides gives

$$V_\varepsilon^\infty(A_i x) \leq \beta_\varepsilon^2 \left(V_\varepsilon^\infty(x) - \|x\|_2^2\right) \leq \beta_\varepsilon^2 V_\varepsilon^\infty(x).$$

Taking square roots yields

$$p_\varepsilon(A_i x) \leq \beta_\varepsilon p_\varepsilon(x).$$

Proof of (7). From (6),

$$\|A_i\|_{p_\varepsilon} = \sup_{x\in\mathbb{R}^m,\, x\neq 0} \frac{p_\varepsilon(A_i x)}{p_\varepsilon(x)} \leq \beta_\varepsilon.$$

Hence, for any word $\sigma = (\sigma_1, \dots, \sigma_k)$,

$$\|A_\sigma\|_{p_\varepsilon} \leq \prod_{j=1}^{k} \|A_{\sigma_j}\|_{p_\varepsilon} \leq \beta_\varepsilon^k.$$

Taking the maximum over all words of length k, then the kth root, and finally the limit as $k \to \infty$, gives

$$\rho(\mathcal{H}) \leq \beta_\varepsilon.$$

This completes the proof. □

For the direct Q-learning family

$$\mathcal{M}_\alpha = \{M_\pi : \pi \in \Theta\}, \qquad M_\pi = I - \alpha D + \alpha\gamma DP\Pi^\pi,$$

fix $\varepsilon > 0$ such that

$$\beta_\varepsilon := \rho(\mathcal{M}_\alpha) + \varepsilon < 1.$$

Since Θ is finite, enumerate $\mathcal{M}_\alpha$ as $\{M_1, \dots, M_M\}$, where $M = |\Theta|$. For each integer $t \geq 0$, define

$$V_\varepsilon^t(x) := \sum_{\ell=0}^{t} \beta_\varepsilon^{-2\ell} \max_{\pi_1,\dots,\pi_\ell \in \Theta} \|M_{\pi_\ell} \cdots M_{\pi_1} x\|_2^2, \qquad x \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \tag{28}$$

where the term for $\ell = 0$ uses the empty product and is equal to $\|x\|_2^2$. Define

$$V_\varepsilon^\infty(x) := \lim_{t\to\infty} V_\varepsilon^t(x). \tag{29}$$

By Lemma 2, the limit exists and there exists a constant $C_\varepsilon \geq 1$ such that

$$\|x\|_2^2 \leq V_\varepsilon^\infty(x) \leq C_\varepsilon \|x\|_2^2, \qquad \forall x \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}. \tag{30}$$

The following lemma specializes the preceding construction to $\mathcal{M}_\alpha$ and gives the norm equivalence and contraction property used below.

Lemma 3. *With the definitions in Equations (28) and (29), the norm-equivalence statement in Equation (30) holds. In particular,*

$$p_\varepsilon(x) := \sqrt{V_\varepsilon^\infty(x)}$$

is a norm on $\mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$. Moreover, for every deterministic policy $\pi \in \Theta$,

$$V_\varepsilon^\infty(M_\pi x) \leq \beta_\varepsilon^2 \left(V_\varepsilon^\infty(x) - \|x\|_2^2\right), \qquad \forall x \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}. \tag{31}$$

Equivalently,

$$p_\varepsilon(M_\pi x) \leq \beta_\varepsilon p_\varepsilon(x), \qquad \forall x \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \quad \forall \pi \in \Theta.$$

Proof. The state vector and direct-family modes satisfy

$$x \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \qquad M_\pi \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

For the direct-family statement, apply the generic construction to the enumerated family $\mathcal{M}_\alpha = \{M_1, \dots, M_M\}$. The finite-horizon inequality in item 1 gives

$$V_\varepsilon^{t+1}(x) \geq \|x\|_2^2 + \beta_\varepsilon^{-2} \max_{\pi\in\Theta} V_\varepsilon^t(M_\pi x).$$

Therefore, for each fixed π,

$$V_\varepsilon^t(M_\pi x) \leq \beta_\varepsilon^2 \left(V_\varepsilon^{t+1}(x) - \|x\|_2^2\right).$$

Letting $t \to \infty$ gives Equation (31). Taking square roots gives the stated norm contraction.

This completes the proof. □

The next lemma extends the deterministic-mode inequality in Lemma 3 to every stochastic-policy mode.

Lemma 4. *For every stochastic policy μ,*

$$V_\varepsilon^\infty(M_\mu x) \leq \beta_\varepsilon^2 \left(V_\varepsilon^\infty(x) - \|x\|_2^2\right), \qquad \forall x \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}. \tag{32}$$

Proof. The vectors and mode matrices below satisfy

$$x \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \qquad M_\mu, M_\pi \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

For finite t, each summand in Equation (28) is the maximum of convex quadratic functions, and a nonnegative sum of convex functions is convex. Hence V_ε^t is convex for every t. Since

$$V_\varepsilon^\infty(x) = \sup_{t\geq 0} V_\varepsilon^t(x),$$

the function V_ε^∞ is convex as well. Next, fix a stochastic policy μ. By Lemma 7, we have

$$M_\mu = \sum_{\pi\in\Theta} c_\pi(\mu) M_\pi, \qquad c_\pi(\mu) \geq 0, \qquad \sum_{\pi\in\Theta} c_\pi(\mu) = 1.$$

Applying Jensen's inequality to the convex function V_ε^∞, we have

$$V_\varepsilon^\infty(M_\mu x) = V_\varepsilon^\infty\left(\sum_{\pi\in\Theta} c_\pi(\mu) M_\pi x\right) \leq \sum_{\pi\in\Theta} c_\pi(\mu) V_\varepsilon^\infty(M_\pi x) \leq \max_{\pi\in\Theta} V_\varepsilon^\infty(M_\pi x).$$

Using Equation (31),

$$\max_{\pi\in\Theta} V_\varepsilon^\infty(M_\pi x) \leq \beta_\varepsilon^2 \left(V_\varepsilon^\infty(x) - \|x\|_2^2\right),$$

which proves Equation (32). □

C Deterministic finite-time error bound

The SLS representation in Theorem 1 makes it possible to analyze the finite-time behavior of the noise-free Q-learning recursion through SLS theory and the JSR (Liberzon, 2003; Lin and Antsaklis, 2009; Shorten et al., 2007; Rota and Strang, 1960; Tsitsiklis and Blondel, 1997; Jungers, 2009). To state the bound, fix $\varepsilon > 0$ and write

$$\beta_\varepsilon := \rho(\mathcal{M}_\alpha) + \varepsilon.$$

For each integer $t \geq 0$, define

$$V_\varepsilon^t(x) := \sum_{\ell=0}^{t} \beta_\varepsilon^{-2\ell} \max_{\pi_1,\ldots,\pi_\ell\in\Theta} \|M_{\pi_\ell}\cdots M_{\pi_1} x\|_2^2, \qquad x \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|},$$

with the empty-product term equal to $\|x\|_2^2$, and define

$$V_\varepsilon^\infty(x) := \lim_{t\to\infty} V_\varepsilon^t(x).$$

The construction is established in Section B. When $\beta_\varepsilon < 1$, Lemma 3 shows that V_ε^∞ is finite and satisfies

$$\|x\|_2^2 \leq V_\varepsilon^\infty(x) \leq C_\varepsilon \|x\|_2^2, \qquad \forall x \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|},$$

for some $C_\varepsilon \geq 1$. The following theorem gives the deterministic version of the finite-time result.

Theorem 6. *Let Assumption 1 hold. Fix any $\varepsilon > 0$ such that $\beta_\varepsilon < 1$, and let V_ε^∞ and C_ε be defined above. Consider the deterministic, noise-free recursion*

$$Q_{k+1} = Q_k + \alpha D\big(F(Q_k) - Q_k\big), \qquad k \geq 0. \tag{33}$$

Then, for every $k \geq 0$,

$$V_\varepsilon^\infty(Q_k - Q^*) \leq \beta_\varepsilon^{2k} V_\varepsilon^\infty(Q_0 - Q^*). \tag{34}$$

Consequently,

$$\|Q_k - Q^*\|_\infty \leq \sqrt{C_\varepsilon}\, \beta_\varepsilon^k \|Q_0 - Q^*\|_2. \tag{35}$$

Proof. The error vectors and mode matrices satisfy

$$Q_k - Q^*, Q_0 - Q^* \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \qquad M_{\mu_k} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

For the deterministic recursion in Equation (33), the state-wise linearization in Lemma 6 yields a stochastic policy μ_k such that

$$Q_{k+1} - Q^* = M_{\mu_k}(Q_k - Q^*).$$

By Lemma 7, $M_{\mu_k} \in \mathrm{co}(\mathcal{M}_\alpha)$. Therefore, Lemma 4 gives

$$\begin{aligned} V_\varepsilon^\infty(Q_{k+1} - Q^*) &= V_\varepsilon^\infty\big(M_{\mu_k}(Q_k - Q^*)\big) \\ &\le \beta_\varepsilon^2 \left(V_\varepsilon^\infty(Q_k - Q^*) - \|Q_k - Q^*\|_2^2\right) \\ &\le \beta_\varepsilon^2 V_\varepsilon^\infty(Q_k - Q^*). \end{aligned}$$

Iteration proves Equation (34). Finally, the norm-equivalence estimate in Equation (30) implies

$$\begin{aligned} \|Q_k - Q^*\|_\infty &\le \|Q_k - Q^*\|_2 \\ &\le \sqrt{V_\varepsilon^\infty(Q_k - Q^*)} \\ &\le \beta_\varepsilon^k \sqrt{V_\varepsilon^\infty(Q_0 - Q^*)} \\ &\le \sqrt{C_\varepsilon}\, \beta_\varepsilon^k \|Q_0 - Q^*\|_2, \end{aligned}$$

which is Equation (35). □

This theorem isolates the deterministic drift rate of the direct SLS representation.

D STOCHASTIC FINITE-TIME ERROR BOUND

The following proposition extends the product-defined Lyapunov estimate to the stochastic recursion and separates the transient from the noise floor.

Proposition 4. *Let Assumption 1 hold and consider the i.i.d. Q-learning recursion in Section 2.4. Fix any $\varepsilon > 0$ such that $\beta_\varepsilon < 1$, and let V_ε^∞ and C_ε be defined above. Then, for every $k \ge 0$,*

$$\begin{aligned} \mathbb{E}[V_\varepsilon^\infty(Q_k - Q^*)] \le \beta_\varepsilon^{2k} V_\varepsilon^\infty(Q_0 - Q^*) \\ + \frac{\alpha^2 C_\varepsilon^2 W_{\max}}{1 - \beta_\varepsilon^2}\left(1 - \beta_\varepsilon^{2k}\right). \end{aligned} \tag{36}$$

Consequently,

$$\mathbb{E}[\|Q_k - Q^*\|_\infty] \le \sqrt{C_\varepsilon}\, \beta_\varepsilon^k \|Q_0 - Q^*\|_2 + \alpha C_\varepsilon \sqrt{\frac{W_{\max}}{1 - \beta_\varepsilon^2}}. \tag{37}$$

The constant C_ε is instance dependent and is not claimed to be uniformly controlled as $\alpha, \varepsilon \downarrow 0$; this is an existence result for the product-defined Lyapunov function.

Proof. The error and noise vectors and the active mode satisfy

$$Q_k - Q^*, Q_{k+1} - Q^*, w_k \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \qquad M_{\mu_k} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

By Theorem 1,

$$Q_{k+1} - Q^* = M_{\mu_k}(Q_k - Q^*) + \alpha w_k.$$

Because $Q_k - Q^*$ and μ_k are $\mathcal{F}_k$-measurable, $M_{\mu_k}(Q_k - Q^*)$ is also $\mathcal{F}_k$-measurable. By Lemma 3,

$$p_\varepsilon(x) := \sqrt{V_\varepsilon^\infty(x)}$$

is a norm. The proof therefore uses the triangle inequality rather than a cross-term cancellation. Specifically,

$$p_\varepsilon(Q_{k+1} - Q^*) \le p_\varepsilon(M_{\mu_k}(Q_k - Q^*)) + \alpha p_\varepsilon(w_k).$$

Squaring both sides gives

$$\begin{aligned} V_\varepsilon^\infty(Q_{k+1} - Q^*) \le V_\varepsilon^\infty(M_{\mu_k}(Q_k - Q^*)) \\ + 2\alpha p_\varepsilon(M_{\mu_k}(Q_k - Q^*))\, p_\varepsilon(w_k) + \alpha^2 V_\varepsilon^\infty(w_k). \end{aligned} \tag{38}$$

We next bound the three terms on the right-hand side of Equation (38). Since μ_k is a stochastic policy, Lemma 4 gives

$$V_\varepsilon^\infty(M_{\mu_k}(Q_k - Q^*)) \le \beta_\varepsilon^2 \left(V_\varepsilon^\infty(Q_k - Q^*) - \|Q_k - Q^*\|_2^2\right). \tag{39}$$

Taking square roots in Equation (39) yields

$$p_\varepsilon(M_{\mu_k}(Q_k - Q^*)) \le \beta_\varepsilon \sqrt{V_\varepsilon^\infty(Q_k - Q^*) - \|Q_k - Q^*\|_2^2}. \tag{40}$$

The norm equivalence in Equation (30) also implies

$$V_\varepsilon^\infty(Q_k - Q^*) - \|Q_k - Q^*\|_2^2 \le (C_\varepsilon - 1)\|Q_k - Q^*\|_2^2. \tag{41}$$

Combining Equations (40) and (41), we obtain

$$p_\varepsilon(M_{\mu_k}(Q_k - Q^*)) \le \beta_\varepsilon \sqrt{C_\varepsilon - 1}\, \|Q_k - Q^*\|_2. \tag{42}$$

For the noise term, conditional Jensen's inequality and Equation (30) and Lemma 5 give

$$\begin{aligned} \mathbb{E}[p_\varepsilon(w_k) \mid \mathcal{F}_k] &\le \sqrt{\mathbb{E}[p_\varepsilon(w_k)^2 \mid \mathcal{F}_k]} \\ &= \sqrt{\mathbb{E}[V_\varepsilon^\infty(w_k) \mid \mathcal{F}_k]} \\ &\le \sqrt{C_\varepsilon \mathbb{E}[\|w_k\|_2^2 \mid \mathcal{F}_k]} \le \sqrt{C_\varepsilon W_{\max}}. \end{aligned} \tag{43}$$

Similarly,

$$\mathbb{E}[V_\varepsilon^\infty(w_k) \mid \mathcal{F}_k] \le C_\varepsilon \mathbb{E}[\|w_k\|_2^2 \mid \mathcal{F}_k] \le C_\varepsilon W_{\max}. \tag{44}$$

Substituting Equations (39) and (42) to (44) into Equation (38), taking conditional expectation, and using the $\mathcal{F}_k$-measurability of $p_\varepsilon(M_{\mu_k}(Q_k - Q^*))$, we have

$$\begin{aligned} &\mathbb{E}[V_\varepsilon^\infty(Q_{k+1} - Q^*) \mid \mathcal{F}_k] \\ &\quad \le \beta_\varepsilon^2 V_\varepsilon^\infty(Q_k - Q^*) - \beta_\varepsilon^2 \|Q_k - Q^*\|_2^2 \\ &\qquad + 2\alpha\beta_\varepsilon \sqrt{C_\varepsilon(C_\varepsilon - 1)W_{\max}}\, \|Q_k - Q^*\|_2 + \alpha^2 C_\varepsilon W_{\max}. \end{aligned} \tag{45}$$

Apply $2ab \le a^2 + b^2$ with

$$a = \beta_\varepsilon \|Q_k - Q^*\|_2, \qquad b = \alpha\sqrt{C_\varepsilon(C_\varepsilon - 1)W_{\max}}.$$

Then

$$\begin{aligned} &2\alpha\beta_\varepsilon \sqrt{C_\varepsilon(C_\varepsilon - 1)W_{\max}}\, \|Q_k - Q^*\|_2 \\ &\quad \le \beta_\varepsilon^2 \|Q_k - Q^*\|_2^2 + \alpha^2 C_\varepsilon(C_\varepsilon - 1)W_{\max}. \end{aligned} \tag{46}$$

The first term on the right-hand side of Equation (46) cancels the negative Euclidean-norm term in Equation (45). Hence

$$\mathbb{E}[V_\varepsilon^\infty(Q_{k+1} - Q^*) \mid \mathcal{F}_k] \le \beta_\varepsilon^2 V_\varepsilon^\infty(Q_k - Q^*) + \alpha^2 C_\varepsilon^2 W_{\max}. \tag{47}$$

Taking total expectation and setting $a_k := \mathbb{E}[V_\varepsilon^\infty(Q_k - Q^*)]$ gives

$$a_{k+1} \le \beta_\varepsilon^2 a_k + \alpha^2 C_\varepsilon^2 W_{\max}.$$

Iteration yields

$$a_k \le \beta_\varepsilon^{2k} a_0 + \alpha^2 C_\varepsilon^2 W_{\max} \sum_{j=0}^{k-1} \beta_\varepsilon^{2j}. \tag{48}$$

Since

$$\sum_{j=0}^{k-1} \beta_\varepsilon^{2j} = \frac{1-\beta_\varepsilon^{2k}}{1-\beta_\varepsilon^2}, \tag{49}$$

combining Equations (48) and (49) and using the deterministic initial condition proves Equation (36).

Finally, the lower bound in Equation (30) implies

$$\|Q_k - Q^*\|_\infty \le \|Q_k - Q^*\|_2 \le \sqrt{V_\varepsilon^\infty(Q_k - Q^*)}.$$

Therefore, Jensen's inequality and Equation (36) give

$$\begin{aligned}
\mathbb{E}[\|Q_k - Q^*\|_\infty] &\le \sqrt{\mathbb{E}[V_\varepsilon^\infty(Q_k - Q^*)]} \\
&\le \sqrt{C_\varepsilon \beta_\varepsilon^{2k} \|Q_0 - Q^*\|_2^2 + \frac{\alpha^2 C_\varepsilon^2 W_{\max}}{1-\beta_\varepsilon^2}} \\
&\le \sqrt{C_\varepsilon}\, \beta_\varepsilon^k \|Q_0 - Q^*\|_2 + \alpha C_\varepsilon \sqrt{\frac{W_{\max}}{1-\beta_\varepsilon^2}},
\end{aligned}$$

where the second inequality also uses $1 - \beta_\varepsilon^{2k} \le 1$, and the last inequality uses $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. This proves Equation (37). □

E Detailed numerical example

The following example gives the complete matrix construction and exact JSR calculation summarized in Section 8.

Example 1. *Use the state-action ordering* $(1,1),(2,1),(1,2),(2,2)$ *and consider an i.i.d. asynchronous Q-learning problem with* $\mathcal{S} = \{1,2\}$, $\mathcal{A} = \{1,2\}$, $\gamma = 0.9$,

$$d = (0.10, 0.40, 0.10, 0.40), \qquad P(1 \mid s,a) = P(2 \mid s,a) = \frac{1}{2}$$

for every $s \in \mathcal{S}$ *and* $a \in \mathcal{A}$. *Define the four deterministic policies by*

$$\begin{aligned}
&\pi_1(1) = 1, \quad \pi_1(2) = 1, \qquad \pi_2(1) = 1, \quad \pi_2(2) = 2, \\
&\pi_3(1) = 2, \quad \pi_3(2) = 1, \qquad \pi_4(1) = 2, \quad \pi_4(2) = 2.
\end{aligned}$$

The associated matrices are

$$M_{\pi_1} = \begin{bmatrix} 1-\frac{11\alpha}{200} & \frac{9\alpha}{200} & 0 & 0 \\ \frac{9\alpha}{50} & 1-\frac{11\alpha}{50} & 0 & 0 \\ \frac{9\alpha}{200} & \frac{9\alpha}{200} & 1-\frac{\alpha}{10} & 0 \\ \frac{9\alpha}{50} & \frac{9\alpha}{50} & 0 & 1-\frac{2\alpha}{5} \end{bmatrix}.$$

$$M_{\pi_2} = \begin{bmatrix} 1-\frac{11\alpha}{200} & 0 & 0 & \frac{9\alpha}{200} \\ \frac{9\alpha}{50} & 1-\frac{2\alpha}{5} & 0 & \frac{9\alpha}{50} \\ \frac{9\alpha}{200} & 0 & 1-\frac{\alpha}{10} & \frac{9\alpha}{200} \\ \frac{9\alpha}{50} & 0 & 0 & 1-\frac{11\alpha}{50} \end{bmatrix}.$$

$$M_{\pi_3} = \begin{bmatrix} 1-\frac{\alpha}{10} & \frac{9\alpha}{200} & \frac{9\alpha}{200} & 0 \\ 0 & 1-\frac{11\alpha}{50} & \frac{9\alpha}{50} & 0 \\ 0 & \frac{9\alpha}{200} & 1-\frac{11\alpha}{200} & 0 \\ 0 & \frac{9\alpha}{50} & \frac{9\alpha}{50} & 1-\frac{2\alpha}{5} \end{bmatrix}.$$

$$M_{\pi_4} = \begin{bmatrix} 1-\frac{\alpha}{10} & 0 & \frac{9\alpha}{200} & \frac{9\alpha}{200} \\ 0 & 1-\frac{2\alpha}{5} & \frac{9\alpha}{50} & \frac{9\alpha}{50} \\ 0 & 0 & 1-\frac{11\alpha}{200} & \frac{9\alpha}{200} \\ 0 & 0 & \frac{9\alpha}{50} & 1-\frac{11\alpha}{50} \end{bmatrix}.$$

Set

$$\kappa := \frac{11+\sqrt{265}}{24}, \qquad q := [\kappa \quad 1 \quad \kappa \quad 1]^\top, \qquad v := [\kappa \quad 1]^\top.$$

The two action coordinates agree at each state, so for every $\pi \in \Theta$,

$$\Pi^\pi q = v, \qquad Pv = \frac{\kappa+1}{2}\mathbf{1}_4.$$

Consequently,

$$\begin{aligned} M_\pi q &= (I-\alpha D)q + \frac{9\alpha}{10}DPv \\ &= \begin{bmatrix} \kappa + \dfrac{\alpha(9-11\kappa)}{200} \\ 1 + \dfrac{\alpha(9\kappa-11)}{50} \\ \kappa + \dfrac{\alpha(9-11\kappa)}{200} \\ 1 + \dfrac{\alpha(9\kappa-11)}{50} \end{bmatrix}. \end{aligned}$$

The selected value of κ is the positive root of

$$12\kappa^2 - 11\kappa - 3 = 0. \tag{1}$$

The first equality below follows directly from Equation (1). Indeed,

$$\begin{aligned} \frac{11\kappa-9}{200\kappa} = \frac{11-9\kappa}{50} &\iff 50(11\kappa-9) = 200\kappa(11-9\kappa) \\ &\iff 11\kappa - 9 = 4\kappa(11-9\kappa) \\ &\iff 36\kappa^2 - 33\kappa - 9 = 0 \\ &\iff 3(12\kappa^2 - 11\kappa - 3) = 0, \end{aligned} \tag{2}$$

which is exactly Equation (1). The closed form of κ gives

$$\begin{aligned} \frac{11-9\kappa}{50} &= \frac{1}{50}\left(11 - 9\frac{11+\sqrt{265}}{24}\right) \\ &= \frac{1}{50}\frac{165-9\sqrt{265}}{24} \\ &= \frac{55-3\sqrt{265}}{400}. \end{aligned} \tag{3}$$

Combining Equations (2) and (3) gives

$$\frac{11\kappa-9}{200\kappa} = \frac{11-9\kappa}{50} = \frac{55-3\sqrt{265}}{400}. \tag{4}$$

Define

$$\beta := 1 - \frac{55-3\sqrt{265}}{400}\alpha. \tag{5}$$

Using Equation (4), the first and third coordinates of $M_\pi q$ satisfy

$$\begin{aligned} \kappa + \frac{\alpha(9-11\kappa)}{200} &= \kappa\left(1 - \frac{11\kappa-9}{200\kappa}\alpha\right) \\ &= \kappa\left(1 - \frac{55-3\sqrt{265}}{400}\alpha\right) = \beta\kappa, \end{aligned} \tag{6}$$

and the second and fourth coordinates satisfy

$$\begin{aligned} 1 + \frac{\alpha(9\kappa-11)}{50} &= 1 - \frac{11-9\kappa}{50}\alpha \\ &= 1 - \frac{55-3\sqrt{265}}{400}\alpha = \beta. \end{aligned} \tag{7}$$

Substituting Equations (6) and (7) into the vector expression for $M_\pi q$ yields, for every $\pi \in \Theta$,

$$M_\pi q = \beta q = \left(1 - \frac{55 - 3\sqrt{265}}{400}\alpha\right) q. \tag{8}$$

Moreover, $\sqrt{265} > 13$ implies $\kappa > 1$, and

$$q \geq \mathbf{1}_4, \qquad q = \mathbf{1}_2 \otimes v.$$

Thus (β, q, v) satisfies Equation (8) with equality in Equations (8b) and (8c), where β is defined in Equation (5). Since $55 - 3\sqrt{265} > 0$, we have $\beta < 1$. Therefore, Theorem 4 gives

$$\rho(\mathcal{M}_\alpha) \leq \beta. \tag{9}$$

For each $j \in \{1, 2, 3, 4\}$, Equation (8) gives $M_{\pi_j} q = \beta q$. Because $q \neq 0$ and $\beta > 0$, β is an eigenvalue of every M_{π_j}, and hence

$$\rho(M_{\pi_j}) \geq \beta, \qquad j \in \{1, 2, 3, 4\}. \tag{10}$$

Repeating a fixed mode is an admissible switching sequence, so

$$\rho(\mathcal{M}_\alpha) \geq \rho(M_{\pi_j}) \geq \beta.$$

Combining this lower bound with Equation (9) yields $\rho(\mathcal{M}_\alpha) = \beta$. Moreover,

$$\begin{aligned} \frac{55 - 3\sqrt{265}}{400} > \frac{1}{100} &\iff 55 - 3\sqrt{265} > 4 \\ &\iff 17 > \sqrt{265} \\ &\iff 289 > 265. \end{aligned}$$

Therefore,

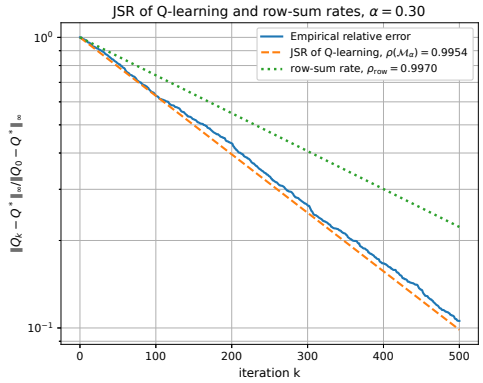
$$\rho(\mathcal{M}_\alpha) = 1 - \frac{55 - 3\sqrt{265}}{400}\alpha < 1 - 0.01\alpha = \rho_{\text{row}}. \tag{11}$$

Thus the values of $\rho(\mathcal{M}_\alpha)$ for $\alpha = 0.30, 0.60, 0.90$ are exactly 0.995377, 0.990755, and 0.986132, respectively, while the corresponding row-sum rates are 0.997000, 0.994000, and 0.991000.

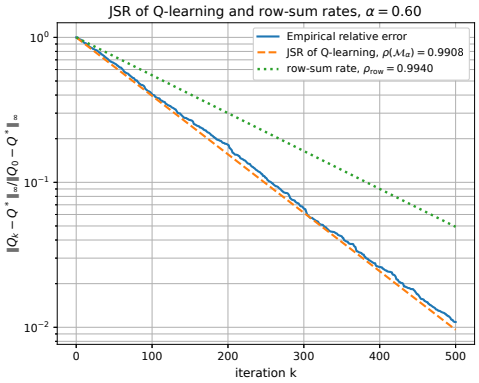
We next simulate Q-learning on the same problem to illustrate the rate separation in Equation (11). Let $r(s, 1, s') = 1$ and $r(s, 2, s') = -9$ for all $s, s' \in \mathcal{S}$. Action 1 is optimal in both states, and $Q^*(s, 1) = 10$ and $Q^*(s, 2) = 0$ for every $s \in \mathcal{S}$. The initial condition is $Q_0 = 0$. For each step-size, the empirical curve averages 20 independent trajectories and plots

$$\frac{\|Q_k - Q^*\|_\infty}{\|Q_0 - Q^*\|_\infty}.$$

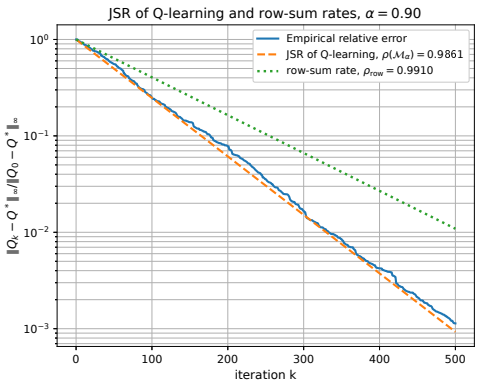
The simulation uses the reward $r(s_k, a_k, s_{k+1})$ without an additional reward perturbation. Each figure compares the averaged relative error with the normalized curves based on the exact JSR of Q-learning and the row-sum rate. These curves illustrate the exact rate separation in Equation (11); they are not empirical sample-complexity bounds.



(a) $\alpha = 0.30$



(b) $\alpha = 0.60$



(c) $\alpha = 0.90$

Figure 2: Relative Q-learning error for three step-sizes, averaged over 20 independent seeds. Each panel compares the empirical relative error with the curves based on the exact JSR of Q-learning and the row-sum rate, both normalized to start at one.

F Optimal LP value: feasibility, nonattainment, and certificates

We begin by naming the optimal feasibility value of Equation (8), which organizes the attainment and certificate results in this section.

Definition 3. *Define the optimal LP feasibility value by*

$$\beta_\star := \inf\{\beta \in \mathbb{R} : \textit{there exist } q \textit{ and } v \textit{ satisfying Equation (8)}\}.$$

By Theorem 4, $\beta_\star = \rho(\mathcal{M}_\alpha)$.

The following example shows that the infimum can fail to be attained for a general transition kernel.

Example 2. *The infimum in Theorem 4 need not be attained. Consider $\mathcal{S} = \{1, 2\}$, $\mathcal{A} = \{1\}$,*

$$P = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}, \qquad d(1,1) = \frac{1-\gamma}{2-\gamma}, \qquad d(2,1) = \frac{1}{2-\gamma}.$$

The unique deterministic policy produces

$$M_\pi = \begin{bmatrix} 1 - \dfrac{\alpha(1-\gamma)}{2-\gamma} & \dfrac{\alpha\gamma(1-\gamma)}{2-\gamma} \\ 0 & 1 - \dfrac{\alpha(1-\gamma)}{2-\gamma} \end{bmatrix}.$$

Because this matrix is upper triangular,

$$\beta_\star = \rho(\mathcal{M}_\alpha) = \rho(M_\pi) = 1 - \frac{\alpha(1-\gamma)}{2-\gamma}.$$

There is only one action, so one may set $v(s) = q(s,1)$ when constructing a feasible point. To show infeasibility at $\beta = \beta_\star$, however, it is enough to retain the inequality $q(s,1) \le v(s)$ from Equation (8b). The first coordinate of Equation (8c) would require

$$\left(1 - \frac{\alpha(1-\gamma)}{2-\gamma}\right) q(1,1) + \frac{\alpha\gamma(1-\gamma)}{2-\gamma} v(2) \le \left(1 - \frac{\alpha(1-\gamma)}{2-\gamma}\right) q(1,1).$$

After canceling the common term, this becomes

$$\frac{\alpha\gamma(1-\gamma)}{2-\gamma} v(2) \le 0.$$

On the other hand, Equations (8a) and (8b) imply

$$v(2) \ge q(2,1) \ge 1,$$

and $\alpha\gamma(1-\gamma)/(2-\gamma) > 0$. This is a contradiction.

For every fixed $\beta > \beta_\star$, feasibility can be verified explicitly. Set

$$\begin{aligned} q(2,1) &:= 1, \\ q(1,1) &:= \max\left\{1, \frac{\alpha\gamma(1-\gamma)/(2-\gamma)}{\beta - (1 - \alpha(1-\gamma)/(2-\gamma))}\right\}, \\ v(1) &:= q(1,1), \qquad v(2) := q(2,1). \end{aligned}$$

The first two LP inequalities hold by construction. For the dynamic inequality,

$$\begin{aligned} \left(1 - \frac{\alpha(1-\gamma)}{2-\gamma}\right) q(1,1) + \frac{\alpha\gamma(1-\gamma)}{2-\gamma} q(2,1) &\le \beta q(1,1), \\ \left(1 - \frac{\alpha(1-\gamma)}{2-\gamma}\right) q(2,1) &\le \beta q(2,1), \end{aligned}$$

where the first inequality follows from the selected ratio and the second from $\beta > \beta_\star$. Hence the LP value is an infimum rather than a minimum in general. Moreover,

$$\kappa_q = \frac{\max\{q(1,1), q(2,1)\}}{\min\{q(1,1), q(2,1)\}} = q(1,1) \longrightarrow \infty \qquad \text{as } \beta \downarrow \beta_\star.$$

The matrix powers make the same obstruction explicit:

$$M_\pi^k = \begin{bmatrix} \left(1 - \frac{\alpha(1-\gamma)}{2-\gamma}\right)^k & \frac{k\alpha\gamma(1-\gamma)}{2-\gamma}\left(1 - \frac{\alpha(1-\gamma)}{2-\gamma}\right)^{k-1} \\ 0 & \left(1 - \frac{\alpha(1-\gamma)}{2-\gamma}\right)^k \end{bmatrix}.$$

Since M_π^k is nonnegative,

$$\begin{aligned} \frac{\|M_\pi^k\|_\infty}{\beta_\star^k} &= 1 + \frac{k\alpha\gamma(1-\gamma)}{(2-\gamma)\,(1 - \alpha(1-\gamma)/(2-\gamma))} \\ &\longrightarrow \infty. \end{aligned}$$

Thus no finite product prefactor yields a uniform bound at $\beta = \beta_\star$.

The preceding example shows why feasibility at the infimum must be treated separately. The next lemma gives a sufficient condition for attainment and states the limitation for a general transition kernel.

Lemma 8. *Let $\beta_\star$ be defined in Definition 3. If P is entrywise positive, then Equation (8) is feasible at $\beta = \beta_\star$. For a general transition kernel, feasibility at $\beta = \beta_\star$ need not hold.*

Proof. The vectors and matrices used below satisfy

$$\bar{Q}, q \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \quad v \in \mathbb{R}^{|\mathcal{S}|}, \quad D \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}, \quad P \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times|\mathcal{S}|}.$$

Suppose first that P is entrywise positive. The proof of Theorem 3 constructs $\bar{Q} \gg 0$ satisfying

$$\mathcal{T}(\bar{Q}) = \rho(\mathcal{M}_\alpha)\bar{Q} = \beta_\star \bar{Q}.$$

Set

$$q := \frac{\bar{Q}}{\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} \bar{Q}(s,a)}, \qquad v(s) := \max_{a\in\mathcal{A}} q(s,a).$$

For every $(s,a) \in \mathcal{S}\times\mathcal{A}$, the normalization of $\bar{Q}$ gives

$$q(s,a) = \frac{\bar{Q}(s,a)}{\min_{(\bar{s},\bar{a})\in\mathcal{S}\times\mathcal{A}} \bar{Q}(\bar{s},\bar{a})} \geq 1, \tag{12}$$

and the definition of v gives

$$q(s,a) \leq \max_{u\in\mathcal{A}} q(s,u) = v(s). \tag{13}$$

Equivalently, Equations (12) and (13) can be written in vector form as

$$q \geq \mathbf{1}_{|\mathcal{S}|\,|\mathcal{A}|}, \qquad q \leq \mathbf{1}_{|\mathcal{A}|} \otimes v. \tag{14}$$

Positive homogeneity of $\mathcal{T}$ gives

$$\begin{aligned}(I - \alpha D)q + \alpha\gamma DPv &= \mathcal{T}(q)\\ &= \beta_\star q.\end{aligned}$$

Therefore $(\beta_\star, q, v)$ satisfies Equation (8). For a general transition kernel, Example 2 gives an explicit instance for which no feasible (q, v) exists at $\beta = \beta_\star$. □

A feasible LP certificate also yields a concrete common weighted norm; the next proposition states the resulting mode and product bounds.

Proposition 5 (Explicit common weighted norm)**.** *Suppose that (β, q, v) satisfies Equation (8) and $\beta < 1$. Define*

$$\|x\|_{\infty,q} := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{|x(s,a)|}{q(s,a)}. \tag{15}$$

The induced matrix norm associated with Equation (15) is

$$\|A\|_{\infty,q} := \sup_{x\neq 0} \frac{\|Ax\|_{\infty,q}}{\|x\|_{\infty,q}}. \tag{16}$$

Then, for every stochastic policy μ,

$$\|M_\mu x\|_{\infty,q} \leq \beta \|x\|_{\infty,q}. \tag{17}$$

Consequently, for every $k \geq 0$ and every $M_0, \ldots, M_{k-1} \in \mathrm{co}(\mathcal{M}_\alpha)$,

$$\|M_{k-1}\cdots M_0\|_\infty \leq \kappa_q \beta^k, \qquad \kappa_q := \frac{\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} q(s,a)}{\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} q(s,a)}, \tag{18}$$

where the product of length zero is the identity.

Proof. The vectors and matrices used below satisfy

$$q, x \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \quad v, \Pi^\mu q \in \mathbb{R}^{|\mathcal{S}|}, \quad M_\mu, \widetilde{M}, M_i \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

For every stochastic policy μ, Equation (8b) gives, for each state s,

$$(\Pi^\mu q)(s) = \sum_{a\in\mathcal{A}} \mu(a \mid s) q(s,a) \le \sum_{a\in\mathcal{A}} \mu(a \mid s) v(s) = v(s). \tag{19}$$

Thus $\Pi^\mu q \le v$. Using Equation (19) and then Equation (8c), we obtain

$$\begin{aligned} M_\mu q &= (I - \alpha D)q + \alpha\gamma DP\Pi^\mu q \\ &\le (I - \alpha D)q + \alpha\gamma DPv \\ &\le \beta q. \end{aligned} \tag{20}$$

For any vector x, Equation (15) gives, for every $(\bar{s}, \bar{a}) \in \mathcal{S} \times \mathcal{A}$,

$$\frac{|x(\bar{s},\bar{a})|}{q(\bar{s},\bar{a})} \le \|x\|_{\infty,q}, \qquad |x(\bar{s},\bar{a})| \le \|x\|_{\infty,q} q(\bar{s},\bar{a}). \tag{21}$$

Because M_μ is nonnegative, Equations (20) and (21) imply, for every $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned} |(M_\mu x)(s,a)| &\le (M_\mu |x|)(s,a) \\ &= \sum_{(\bar{s},\bar{a})\in\mathcal{S}\times\mathcal{A}} [M_\mu]_{(s,a),(\bar{s},\bar{a})} |x(\bar{s},\bar{a})| \\ &\le \|x\|_{\infty,q} \sum_{(\bar{s},\bar{a})\in\mathcal{S}\times\mathcal{A}} [M_\mu]_{(s,a),(\bar{s},\bar{a})} q(\bar{s},\bar{a}) \\ &= \|x\|_{\infty,q} (M_\mu q)(s,a) \\ &\le \beta \|x\|_{\infty,q} q(s,a). \end{aligned} \tag{22}$$

Dividing Equation (22) by $q(s,a) > 0$ and taking the maximum over (s,a) gives Equation (17).

For $\widetilde{M} = \sum_{\pi\in\Theta} c_\pi M_\pi \in \mathrm{co}(\mathcal{M}_\alpha)$, convexity of the norm and Equation (17) give

$$\begin{aligned} \|\widetilde{M}x\|_{\infty,q} &\le \sum_{\pi\in\Theta} c_\pi \|M_\pi x\|_{\infty,q} \\ &\le \beta \|x\|_{\infty,q}. \end{aligned} \tag{23}$$

Iterating Equation (23) over arbitrary products gives the vector inequality

$$\|M_{k-1} \cdots M_0 x\|_{\infty,q} \le \beta^k \|x\|_{\infty,q}, \qquad M_i \in \mathrm{co}(\mathcal{M}_\alpha). \tag{24}$$

By the induced-norm definition in Equation (16), Equation (24) is equivalent to

$$\|M_{k-1} \cdots M_0\|_{\infty,q} \le \beta^k, \qquad M_i \in \mathrm{co}(\mathcal{M}_\alpha). \tag{25}$$

The weighted and ordinary infinity norms satisfy

$$\|x\|_\infty \le \left(\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} q(s,a) \right) \|x\|_{\infty,q}, \qquad \|x\|_{\infty,q} \le \frac{\|x\|_\infty}{\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} q(s,a)}. \tag{26}$$

Combining Equations (24) and (26) gives

$$\|M_{k-1} \cdots M_0 x\|_\infty \le \kappa_q \beta^k \|x\|_\infty. \tag{27}$$

Taking the ordinary induced infinity norm in Equation (27) proves Equation (18). □

The next lemma translates the weighted-norm estimate into an explicit bound on the product-growth constant K_β.

Lemma 9. *Suppose that (β, q, v) satisfies Equation (8) and $\beta < 1$. Let K_β denote the product-growth constant in Equation (5) with β in place of β_ε. Then*

$$1 \le K_\beta \le \kappa_q < \infty. \tag{28}$$

Moreover, every product generated by $\mathrm{co}(\mathcal{M}_\alpha)$ satisfies

$$\|M_{\ell-1} \cdots M_0\|_\infty \le \kappa_q \beta^\ell, \qquad \ell \ge 0. \tag{29}$$

By the definition of K_β, the same products also satisfy $\|M_{\ell-1} \cdots M_0\|_\infty \le K_\beta \beta^\ell$. If $\beta \ge \rho_{\mathrm{row}}$, then the all-ones certificate gives $K_\beta = 1$.

Proof. Every factor in the products below satisfies

$$M_0, \dots, M_{\ell-1} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

By Equation (18), every product of length ℓ generated by $\mathrm{co}(\mathcal{M}_\alpha)$ satisfies

$$\beta^{-\ell}\|M_{\ell-1} \cdots M_0\|_\infty \le \kappa_q. \tag{30}$$

Taking the two suprema in the definition of K_β and using Equation (30) gives $K_\beta \le \kappa_q < \infty$. The product of length zero gives $K_\beta \ge 1$, and Equation (29) follows directly from Equation (18). If $\beta \ge \rho_{\mathrm{row}}$, then Equation (17) yields

$$\begin{aligned} \beta^{-\ell}\|M_{\ell-1} \cdots M_0\|_\infty &\le \left(\frac{\rho_{\mathrm{row}}}{\beta}\right)^\ell \\ &\le 1. \end{aligned} \tag{31}$$

Together with the length-zero product, Equation (31) proves $K_\beta = 1$. □

Combining that certificate with the i.i.d. reference-system analysis gives the following explicit finite-time bound.

Corollary 2. *Let Assumption 1 hold, let $W_{\max}$ be defined in Equation (4), and suppose that (β, q, v) satisfies Equation (8) with $\beta < 1$. Then, for every $k \ge 0$,*

$$\begin{aligned} \mathbb{E}[\|Q_k - Q^*\|_\infty] \le |\mathcal{S}|\,|\mathcal{A}| &\sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}}\left(1 + \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)}\right) \\ &+ \kappa_q \beta^k \|Q_0 - Q^*\|_\infty \\ &+ 2\alpha\gamma d_{\max}\kappa_q^2 k\beta^{k-1}\|Q_0 - Q^*\|_\infty, \end{aligned}$$

where $k\beta^{k-1} = 0$ when $k = 0$. In particular, if the infimum in Theorem 4 is attained, the same bound holds with $\beta = \rho(\mathcal{M}_\alpha)$.

Proof. The reference, residual, and noise vectors and the mode matrices satisfy

$$x_k, z_k, y_k, w_k, p_k, u_k, v_k, Q_k - Q^* \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \qquad M_{\bar{\mu}}, M_{\mu_k} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

Fix an arbitrary stochastic policy $\bar{\mu}$ and use the same reference system as in Appendix I:

$$x_{k+1} = M_{\bar{\mu}} x_k + \alpha w_k, \qquad x_0 = Q_0 - Q^*.$$

Write $x_k = z_k + y_k$, where

$$\begin{aligned} z_{k+1} &= M_{\bar{\mu}} z_k, & z_0 &= Q_0 - Q^*, \\ y_{k+1} &= M_{\bar{\mu}} y_k + \alpha w_k, & y_0 &= 0. \end{aligned}$$

Since $M_{\bar{\mu}} \in \mathrm{co}(\mathcal{M}_\alpha)$, Equation (18) gives

$$\|z_k\|_\infty = \|M_{\bar{\mu}}^k (Q_0 - Q^*)\|_\infty \le \kappa_q \beta^k \|Q_0 - Q^*\|_\infty. \tag{32}$$

The covariance calculation for y_k uses only the martingale-difference property of w_k, the bound in Lemma 5, and the common row contraction $\|M_\mu\|_\infty \le \rho_{\mathrm{row}}$. The same covariance calculation yields

$$\mathbb{E}[\|y_k\|_\infty] \le |\mathcal{S}|\,|\mathcal{A}| \sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}}. \tag{33}$$

Now define

$$p_k := Q_k - Q^* - x_k.$$

The common noise cancels pathwise, so

$$p_{k+1} = M_{\mu_k} p_k + (M_{\mu_k} - M_{\bar{\mu}}) x_k, \qquad p_0 = 0.$$

By the definition of the mode matrices,

$$\|M_{\mu_k} - M_{\bar{\mu}}\|_\infty \le 2\alpha\gamma d_{\max}. \tag{34}$$

Split $p_k = u_k + v_k$, where

$$\begin{aligned} u_{k+1} &= M_{\mu_k} u_k + (M_{\mu_k} - M_{\bar{\mu}}) z_k, \quad u_0 = 0, \\ v_{k+1} &= M_{\mu_k} v_k + (M_{\mu_k} - M_{\bar{\mu}}) y_k, \quad v_0 = 0. \end{aligned}$$

Unrolling the first recursion and applying Equations (18), (32) and (34) gives

$$\begin{aligned} \|u_k\|_\infty &\le \sum_{i=0}^{k-1} \kappa_q \beta^{k-1-i} \cdot 2\alpha\gamma d_{\max} \cdot \kappa_q \beta^i \|Q_0 - Q^*\|_\infty \\ &= 2\alpha\gamma d_{\max} \kappa_q^2 k \beta^{k-1} \|Q_0 - Q^*\|_\infty. \end{aligned} \tag{35}$$

For v_k, the row-contraction argument remains unchanged. Using Equations (33) and (34) and $1 - \rho_{\mathrm{row}} = \alpha d_{\min}(1-\gamma)$, we obtain

$$\begin{aligned} \mathbb{E}[\|v_k\|_\infty] &\le \sum_{i=0}^{k-1} \rho_{\mathrm{row}}^{k-1-i} \cdot 2\alpha\gamma d_{\max} \cdot \mathbb{E}[\|y_i\|_\infty] \\ &\le 2\alpha\gamma d_{\max} |\mathcal{S}|\,|\mathcal{A}| \sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}} \sum_{i=0}^{k-1} \rho_{\mathrm{row}}^{k-1-i} \\ &\le \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)} |\mathcal{S}|\,|\mathcal{A}| \sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}}. \end{aligned} \tag{36}$$

Finally,

$$Q_k - Q^* = z_k + y_k + u_k + v_k.$$

The triangle inequality, followed by Equations (32), (33), (35) and (36), gives

$$\begin{aligned} \mathbb{E}[\|Q_k - Q^*\|_\infty] \le |\mathcal{S}|\,|\mathcal{A}| \sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}} \left(1 + \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)}\right) \\ + \kappa_q \beta^k \|Q_0 - Q^*\|_\infty \\ + 2\alpha\gamma d_{\max} \kappa_q^2 k \beta^{k-1} \|Q_0 - Q^*\|_\infty, \end{aligned}$$

which is the claimed estimate. If the infimum in Theorem 4 is attained, choose a feasible pair (q, v) at $\beta = \rho(\mathcal{M}_\alpha)$; the same argument then gives the final statement. □

G Proof of Lemma 1

We first restate the stability lemma and then prove the uniform exponential product estimate.

Restatement (Lemma 1). *Consider the SLS*

$$x_{k+1} = A_{\sigma_k} x_k, \qquad A_{\sigma_k} \in \mathcal{H},$$

where $\mathcal{H} \subset \mathbb{R}^{m\times m}$ is finite. If $\rho(\mathcal{H}) < 1$, then $x_k \to 0$ for every initial condition and every switching sequence. Moreover, the convergence is exponential.

Proof. The state and switching matrices satisfy

$$x_k, x_0 \in \mathbb{R}^m, \qquad A_i, A_{\sigma_i} \in \mathbb{R}^{m\times m}.$$

Choose β such that $\rho(\mathcal{H}) < \beta < 1$. By the definition of the JSR, there exists an integer $K \geq 1$ such that

$$\sup_{A_1,\dots,A_k\in\mathcal{H}} \|A_k \cdots A_1\|^{1/k} \leq \beta, \qquad k \geq K.$$

Equivalently, the above inequality gives

$$\sup_{A_1,\dots,A_k\in\mathcal{H}} \|A_k \cdots A_1\| \leq \beta^k, \qquad k \geq K. \tag{37}$$

For the finitely many remaining product lengths $0 \leq k < K$, enlarge the multiplicative constant to

$$C := \max\left\{1, \max_{0\leq k<K} \beta^{-k} \sup_{A_1,\dots,A_k\in\mathcal{H}} \|A_k \cdots A_1\|\right\}. \tag{38}$$

Then Equations (37) and (38) imply the uniform product estimate

$$\sup_{A_1,\dots,A_k\in\mathcal{H}} \|A_k \cdots A_1\| \leq C\beta^k, \qquad k \geq 0, \tag{39}$$

where the product of length zero is the identity. For an arbitrary switching sequence, repeated substitution in the recursion gives

$$x_k = A_{\sigma_{k-1}} \cdots A_{\sigma_0} x_0, \qquad k \geq 1.$$

Therefore, by the induced-norm inequality and Equation (39),

$$\begin{aligned}
\|x_k\| &= \|A_{\sigma_{k-1}} \cdots A_{\sigma_0} x_0\| \\
&\leq \|A_{\sigma_{k-1}} \cdots A_{\sigma_0}\| \, \|x_0\| \\
&\leq \sup_{A_1,\dots,A_k\in\mathcal{H}} \|A_k \cdots A_1\| \, \|x_0\| \\
&\leq C\beta^k \|x_0\|, \qquad k \geq 1.
\end{aligned}$$

For $k = 0$, the same estimate follows from $C \geq 1$. Hence

$$\|x_k\| \leq C\beta^k \|x_0\|, \qquad k \geq 0. \tag{40}$$

Since $\beta < 1$, Equation (40) proves exponential convergence to zero. □

H Proof of Theorem 1

We restate the direct representation before deriving it from the Bellman linearization.

Restatement (Theorem 1). *Under Assumption 1, for each $k \geq 0$, there exists an $\mathcal{F}_k$-measurable stochastic policy μ_k such that*

$$Q_{k+1} - Q^* = M_{\mu_k}(Q_k - Q^*) + \alpha w_k,$$

where $M_\mu = I - \alpha D + \alpha\gamma DP\Pi^\mu$, $\mathbb{E}[w_k \mid \mathcal{F}_k] = 0$, and consequently

$$\mathbb{E}[Q_{k+1} - Q^* \mid \mathcal{F}_k] = M_{\mu_k}(Q_k - Q^*).$$

Proof. The vectors and matrices used below satisfy

$$Q_k, Q_{k+1}, Q^*, w_k \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \quad V_{Q_k} - V^* \in \mathbb{R}^{|\mathcal{S}|},$$
$$D, M_{\mu_k} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}, \quad P \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times|\mathcal{S}|}, \quad \Pi^{\mu_k} \in \mathbb{R}^{|\mathcal{S}|\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

Start from the vector Q-learning recursion

$$Q_{k+1} = Q_k + \alpha\{D(F(Q_k) - Q_k) + w_k\}.$$

Subtract Q^* from both sides. Since $Q^* = F(Q^*)$, we obtain

$$\begin{aligned}(Q_{k+1} - Q^*) &= Q_{k+1} - Q^* \\ &= Q_k - Q^* + \alpha D(F(Q_k) - Q_k) + \alpha w_k \\ &= (Q_k - Q^*) + \alpha D\{F(Q_k) - F(Q^*) - (Q_k - Q^*)\} + \alpha w_k.\end{aligned}$$

This identity is purely algebraic; no approximation has been made. Next, we use the form of the Bellman optimality operator,

$$F(Q) = R + \gamma P V_Q,$$

which leads to

$$F(Q_k) - F(Q^*) = \gamma P(V_{Q_k} - V^*).$$

Because Q_k is $\mathcal{F}_k$-measurable and the policy in Lemma 6 can be selected measurably as a function of its input, there exists an $\mathcal{F}_k$-measurable stochastic policy

$$\mu_k := \mu_{Q_k}$$

such that

$$V_{Q_k} - V^* = \Pi^{\mu_k}(Q_k - Q^*).$$

Substituting this into the preceding identity gives

$$F(Q_k) - F(Q^*) = \gamma P \Pi^{\mu_k}(Q_k - Q^*).$$

Hence

$$\begin{aligned}(Q_{k+1} - Q^*) &= (Q_k - Q^*) + \alpha D\{\gamma P \Pi^{\mu_k}(Q_k - Q^*) - (Q_k - Q^*)\} + \alpha w_k \\ &= \left(I - \alpha D + \alpha\gamma D P \Pi^{\mu_k}\right)(Q_k - Q^*) + \alpha w_k \\ &= M_{\mu_k}(Q_k - Q^*) + \alpha w_k.\end{aligned}$$

This proves the direct SLS representation. It remains only to identify the conditional mean. By construction in Section 2.4,

$$\mathbb{E}[w_k \mid \mathcal{F}_k] = 0.$$

Since M_{μ_k} and $Q_k - Q^*$ are $\mathcal{F}_k$-measurable, taking conditional expectation in the last recursion yields

$$\mathbb{E}[Q_{k+1} - Q^* \mid \mathcal{F}_k] = M_{\mu_k}(Q_k - Q^*).$$

This completes the proof. □

I Proof of Theorem 2

The proof uses a fixed-policy reference system driven by the same martingale noise as the direct Q-learning error recursion. The required product-growth bound is established first.

Lemma 19. *Fix any $\varepsilon > 0$ such that*

$$\beta_\varepsilon := \rho(\mathcal{M}_\alpha) + \varepsilon < 1.$$

Let K_{β_ε} be defined in Equation (5). Then $K_{\beta_\varepsilon} < \infty$. Moreover, every product generated by matrices in $\mathrm{co}(\mathcal{M}_\alpha)$ satisfies

$$\|M_{\ell-1} \cdots M_0\|_\infty \le K_{\beta_\varepsilon} \beta_\varepsilon^\ell, \qquad \ell \ge 0. \tag{41}$$

Proof. By Lemma 7,

$$\rho(\mathrm{co}(\mathcal{M}_\alpha)) = \rho(\mathcal{M}_\alpha).$$

Since $\beta_\varepsilon > \rho(\mathrm{co}(\mathcal{M}_\alpha))$, choose

$$\eta \in \big(\rho(\mathrm{co}(\mathcal{M}_\alpha)), \beta_\varepsilon\big).$$

For every $\ell \geq 0$, define the worst-case length-ℓ product size

$$a_\ell := \sup_{M_0,\dots,M_{\ell-1} \in \mathrm{co}(\mathcal{M}_\alpha)} \|M_{\ell-1} \cdots M_0\|_\infty, \qquad a_0 := 1,$$

where

$$M_0, \dots, M_{\ell-1} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

The sequence satisfies $a_{\ell+j} \leq a_\ell a_j$ for all $\ell, j \geq 0$, and the JSR identity gives

$$\lim_{\ell\to\infty} a_\ell^{1/\ell} = \rho(\mathrm{co}(\mathcal{M}_\alpha)) < \eta.$$

Thus there exists $L_\eta \geq 1$ such that $a_\ell \leq \eta^\ell$ for all $\ell \geq L_\eta$. Set

$$C_\eta := \max\left\{1, \max_{0\leq j<L_\eta} \eta^{-j} a_j\right\} < \infty.$$

This finite-length correction and the tail estimate together imply

$$a_\ell \leq C_\eta \eta^\ell, \qquad \ell \geq 0.$$

Therefore,

$$\begin{aligned} K_{\beta_\varepsilon} &= \sup_{\ell\geq 0} \beta_\varepsilon^{-\ell} a_\ell \\ &\leq C_\eta \sup_{\ell\geq 0} \left(\frac{\eta}{\beta_\varepsilon}\right)^\ell \\ &= C_\eta < \infty, \end{aligned}$$

because $0 < \eta/\beta_\varepsilon < 1$ and hence

$$\sup_{\ell\geq 0} \left(\frac{\eta}{\beta_\varepsilon}\right)^\ell = 1.$$

The product estimate Equation (41) follows immediately from the definition of K_{β_ε}. □

With the product estimate in place, we now restate and prove the i.i.d. finite-time bound.

Restatement (Theorem 2)**.** *Let Assumption 1 hold, let $W_{\max}$ be defined in Equation (4), and fix $\varepsilon > 0$ such that $\beta_\varepsilon = \rho(\mathcal{M}_\alpha) + \varepsilon < 1$. Then, for every $k \geq 0$,*

$$\begin{aligned} \mathbb{E}[\|Q_k - Q^*\|_\infty] \leq |\mathcal{S}|\,|\mathcal{A}| &\sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}} \left(1 + \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)}\right) \\ &+ K_{\beta_\varepsilon} \beta_\varepsilon^k \|Q_0 - Q^*\|_\infty \\ &+ 2\alpha\gamma d_{\max} K_{\beta_\varepsilon}^2 k \beta_\varepsilon^{k-1} \|Q_0 - Q^*\|_\infty, \end{aligned}$$

with $k\beta_\varepsilon^{k-1} = 0$ when $k = 0$.

Proof. The reference, residual, and noise vectors and the mode matrices satisfy

$$x_k, z_k, y_k, w_k, p_k, u_k, v_k, Q_k - Q^* \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \qquad M_{\bar{\mu}}, M_{\mu_k} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

We split the proof into three steps.

Step 1: introduce a fixed-policy reference system. Fix an arbitrary stochastic policy $\bar{\mu}$ and define

$$x_{k+1} = M_{\bar{\mu}} x_k + \alpha w_k, \qquad x_0 = Q_0 - Q^*.$$

Decompose

$$x_k = z_k + y_k,$$

where

$$z_{k+1} = M_{\bar{\mu}} z_k, \qquad z_0 = Q_0 - Q^*,$$

and

$$y_{k+1} = M_{\bar{\mu}} y_k + \alpha w_k, \qquad y_0 = 0.$$

Since $M_{\bar{\mu}} \in \mathrm{co}(\mathcal{M}_\alpha)$, Equation (41) gives

$$\|z_k\|_\infty \le K_{\beta_\varepsilon} \beta_\varepsilon^k \|Q_0 - Q^*\|_\infty. \tag{42}$$

We next bound the noise-driven term y_k by the row contraction. Let

$$Y_k := \mathbb{E}[y_k y_k^\top] \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

Since y_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[w_k \mid \mathcal{F}_k] = 0$, the martingale cross terms vanish, and hence

$$Y_{k+1} = M_{\bar{\mu}} Y_k M_{\bar{\mu}}^\top + \alpha^2 \mathbb{E}[w_k w_k^\top].$$

By Lemma 5,

$$\lambda_{\max}(\mathbb{E}[w_k w_k^\top]) \le \mathbb{E}[\|w_k\|_2^2] \le W_{\max}.$$

Unrolling the covariance recursion yields

$$Y_k = \alpha^2 \sum_{i=0}^{k-1} M_{\bar{\mu}}^i \mathbb{E}[w_{k-1-i} w_{k-1-i}^\top] (M_{\bar{\mu}}^\top)^i.$$

Therefore,

$$\begin{aligned}
\lambda_{\max}(Y_k) &\le \alpha^2 W_{\max} \sum_{i=0}^{k-1} \|M_{\bar{\mu}}^i\|_2^2 \\
&\le \alpha^2 W_{\max} |\mathcal{S}|\,|\mathcal{A}| \sum_{i=0}^{k-1} \|M_{\bar{\mu}}^i\|_\infty^2 \\
&\le \alpha^2 W_{\max} |\mathcal{S}|\,|\mathcal{A}| \sum_{i=0}^{k-1} \rho_{\mathrm{row}}^{2i} \\
&\le \frac{\alpha^2 W_{\max} |\mathcal{S}|\,|\mathcal{A}|}{1 - \rho_{\mathrm{row}}} = \frac{\alpha W_{\max} |\mathcal{S}|\,|\mathcal{A}|}{d_{\min}(1-\gamma)}.
\end{aligned}$$

Since $Y_k \succeq 0$,

$$\mathrm{tr}(Y_k) \le |\mathcal{S}|\,|\mathcal{A}| \lambda_{\max}(Y_k) \le \frac{\alpha W_{\max} |\mathcal{S}|^2 |\mathcal{A}|^2}{d_{\min}(1-\gamma)}.$$

Consequently,

$$\mathbb{E}[\|y_k\|_\infty] \le \mathbb{E}[\|y_k\|_2] \le \sqrt{\mathbb{E}[\|y_k\|_2^2]} = \sqrt{\mathrm{tr}(Y_k)} \le |\mathcal{S}|\,|\mathcal{A}| \sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}}. \tag{43}$$

Step 2: subtract the reference system from the direct SLS. Define

$$p_k := Q_k - Q^* - x_k.$$

Using the same noise w_k in both systems gives the pathwise cancellation

$$\begin{aligned}
p_{k+1} &= Q_{k+1} - Q^* - x_{k+1} \\
&= M_{\mu_k}(Q_k - Q^*) + \alpha w_k - M_{\bar{\mu}} x_k - \alpha w_k \\
&= M_{\mu_k} p_k + (M_{\mu_k} - M_{\bar{\mu}}) x_k.
\end{aligned}$$

Furthermore,

$$M_{\mu_k} - M_{\bar{\mu}} = \alpha\gamma D P (\Pi^{\mu_k} - \Pi^{\bar{\mu}}).$$

For every row, the ℓ_1-distance between two stochastic action distributions is at most two; hence

$$\|M_{\mu_k} - M_{\bar{\mu}}\|_\infty \le 2\alpha\gamma d_{\max}. \tag{44}$$

Moreover, $p_0 = Q_0 - Q^* - x_0 = 0$.

Now split

$$p_k = u_k + v_k,$$

where

$$u_{k+1} = M_{\mu_k} u_k + (M_{\mu_k} - M_{\bar{\mu}}) z_k, \qquad u_0 = 0,$$

and

$$v_{k+1} = M_{\mu_k} v_k + (M_{\mu_k} - M_{\bar{\mu}}) y_k, \qquad v_0 = 0.$$

The term u_k is driven by the deterministic reference transient, while v_k is driven by the reference noise floor.

Step 3: bound the deterministic and noise-driven residuals. Unrolling u_k gives

$$u_k = \sum_{i=0}^{k-1} M_{\mu_{k-1}} \cdots M_{\mu_{i+1}} (M_{\mu_i} - M_{\bar{\mu}}) z_i,$$

with the empty product equal to the identity. Since each $M_{\mu_j} \in \mathrm{co}(\mathcal{M}_\alpha)$, Equation (41), Equation (42), and Equation (44) imply

$$\begin{aligned}\|u_k\|_\infty &\le \sum_{i=0}^{k-1} K_{\beta_\varepsilon} \beta_\varepsilon^{k-1-i} \cdot 2\alpha\gamma d_{\max} \cdot K_{\beta_\varepsilon} \beta_\varepsilon^{i} \|Q_0 - Q^*\|_\infty \\ &= 2\alpha\gamma d_{\max} K_{\beta_\varepsilon}^2 k \beta_\varepsilon^{k-1} \|Q_0 - Q^*\|_\infty.\end{aligned}$$

Thus

$$\mathbb{E}[\|u_k\|_\infty] \le 2\alpha\gamma d_{\max} K_{\beta_\varepsilon}^2 k \beta_\varepsilon^{k-1} \|Q_0 - Q^*\|_\infty. \tag{45}$$

For v_k, use the row contraction:

$$v_k = \sum_{i=0}^{k-1} M_{\mu_{k-1}} \cdots M_{\mu_{i+1}} (M_{\mu_i} - M_{\bar{\mu}}) y_i.$$

Pathwise,

$$\|M_{\mu_{k-1}} \cdots M_{\mu_{i+1}}\|_\infty \le \rho_{\mathrm{row}}^{k-1-i}.$$

Using Equations (43) and (44), we obtain

$$\begin{aligned}\mathbb{E}[\|v_k\|_\infty] &\le \sum_{i=0}^{k-1} \rho_{\mathrm{row}}^{k-1-i} \cdot 2\alpha\gamma d_{\max} \cdot \mathbb{E}[\|y_i\|_\infty] \\ &\le 2\alpha\gamma d_{\max} |\mathcal{S}|\,|\mathcal{A}| \sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}} \sum_{i=0}^{k-1} \rho_{\mathrm{row}}^{k-1-i} \\ &\le 2\alpha\gamma d_{\max} |\mathcal{S}|\,|\mathcal{A}| \sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}} \frac{1}{1-\rho_{\mathrm{row}}} \\ &= \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)} |\mathcal{S}|\,|\mathcal{A}| \sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}}.\end{aligned}$$

Hence

$$\mathbb{E}[\|v_k\|_\infty] \le \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)} |\mathcal{S}|\,|\mathcal{A}| \sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}}. \tag{46}$$

Finally, since $x_k = z_k + y_k$,

$$Q_k - Q^* = x_k + p_k = z_k + y_k + u_k + v_k.$$

Taking expectations and applying the triangle inequality gives

$$\mathbb{E}[\|Q_k - Q^*\|_\infty] \le \|z_k\|_\infty + \mathbb{E}[\|y_k\|_\infty] + \mathbb{E}[\|u_k\|_\infty] + \mathbb{E}[\|v_k\|_\infty].$$

Substituting Equations (42), (43), (45) and (46) into the preceding estimate yields

$$\begin{aligned}
\mathbb{E}[\|Q_k - Q^*\|_\infty] &\le K_{\beta_\varepsilon}\beta_\varepsilon^k \|Q_0 - Q^*\|_\infty \\
&\quad + |\mathcal{S}|\,|\mathcal{A}|\sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}} \\
&\quad + 2\alpha\gamma d_{\max} K_{\beta_\varepsilon}^2 k\beta_\varepsilon^{k-1}\|Q_0 - Q^*\|_\infty \\
&\quad + \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)}|\mathcal{S}|\,|\mathcal{A}|\sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}} \\
&= \left(1 + \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)}\right)|\mathcal{S}|\,|\mathcal{A}|\sqrt{\frac{\alpha W_{\max}}{d_{\min}(1-\gamma)}} \\
&\quad + K_{\beta_\varepsilon}\beta_\varepsilon^k \|Q_0 - Q^*\|_\infty \\
&\quad + 2\alpha\gamma d_{\max} K_{\beta_\varepsilon}^2 k\beta_\varepsilon^{k-1}\|Q_0 - Q^*\|_\infty,
\end{aligned}$$

which is exactly Equation (6). □

J Proof of Theorem 3

We use the following standard fixed-point result.

Lemma 20 (Brouwer fixed-point theorem)**.** *Let C be a nonempty compact convex subset of a finite-dimensional real vector space. Every continuous map $f : C \to C$ has a fixed point; that is, there exists $x \in C$ such that $f(x) = x$ (Border, 1985).*

The proof also uses the following standard notions for order-preserving homogeneous maps.

Definition 4 (Monotone map)**.** *Let $C \subseteq \mathbb{R}^m$. A map $f : C \to \mathbb{R}^r$ is monotone with respect to the entrywise order if $x \le y$ implies $f(x) \le f(y)$ for all $x, y \in C$.*

Definition 5 (Positively homogeneous map)**.** *Let $C \subseteq \mathbb{R}^m$. A map $f : C \to \mathbb{R}^r$ is positively homogeneous of degree one if*

$$f(cx) = cf(x)$$

for every $x \in C$ and every scalar $c \ge 0$ such that $cx \in C$.

Definition 6 (Self-map)**.** *For a set C, a map $f : C \to C$ is called a self-map of C.*

For $Q \in \mathbb{R}_+^{|\mathcal{S}|\,|\mathcal{A}|}$, define

$$\mathcal{T}(Q) := (I - \alpha D)Q + \alpha\gamma DP \begin{bmatrix} \max_{a\in\mathcal{A}} Q(1,a) \\ \vdots \\ \max_{a\in\mathcal{A}} Q(|\mathcal{S}|,a) \end{bmatrix}.$$

The next two auxiliary lemmas verify the order and scaling properties of $\mathcal{T}$.

Lemma (Monotonicity of $\mathcal{T}$). *For any $Q, Q' \in \mathbb{R}_+^{|\mathcal{S}|\,|\mathcal{A}|}$, if $Q \le Q'$, then $\mathcal{T}(Q) \le \mathcal{T}(Q')$.*

Proof. The vectors and matrices used below satisfy

$$Q, Q', \mathcal{T}(Q), \mathcal{T}(Q') \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \quad D \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}, \quad P \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times|\mathcal{S}|}.$$

For every $s \in \mathcal{S}$, the entrywise inequality $Q \le Q'$ implies

$$\max_{a\in\mathcal{A}} Q(s,a) \le \max_{a\in\mathcal{A}} Q'(s,a).$$

Hence

$$\begin{bmatrix} \max_{a\in\mathcal{A}} Q(1,a) \\ \vdots \\ \max_{a\in\mathcal{A}} Q(|\mathcal{S}|,a) \end{bmatrix} \le \begin{bmatrix} \max_{a\in\mathcal{A}} Q'(1,a) \\ \vdots \\ \max_{a\in\mathcal{A}} Q'(|\mathcal{S}|,a) \end{bmatrix}.$$

Because $d(s,a) \le 1$ and $\alpha \in (0,1)$, every diagonal entry of $I - \alpha D$ is nonnegative. The matrices D and P are also nonnegative, so multiplication by $I - \alpha D$ and by $\alpha\gamma DP$ preserves entrywise inequalities. Therefore,

$$\begin{aligned} \mathcal{T}(Q) &= (I-\alpha D)Q + \alpha\gamma DP \begin{bmatrix} \max_a Q(1,a) \\ \vdots \\ \max_a Q(|\mathcal{S}|,a) \end{bmatrix} \\ &\le (I-\alpha D)Q' + \alpha\gamma DP \begin{bmatrix} \max_a Q'(1,a) \\ \vdots \\ \max_a Q'(|\mathcal{S}|,a) \end{bmatrix} = \mathcal{T}(Q'). \end{aligned}$$

□

Lemma (Positive homogeneity of $\mathcal{T}$). *For every $Q \in \mathbb{R}_+^{|\mathcal{S}|\,|\mathcal{A}|}$ and every scalar $c \ge 0$, $\mathcal{T}(cQ) = c\mathcal{T}(Q)$.*

Proof. The vectors and matrices used below satisfy

$$Q, \mathcal{T}(Q), \mathcal{T}(cQ) \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \quad D \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}, \quad P \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times|\mathcal{S}|}.$$

For each $s \in \mathcal{S}$ and $c \ge 0$,

$$\max_{a\in\mathcal{A}} \big(cQ(s,a)\big) = c \max_{a\in\mathcal{A}} Q(s,a).$$

Substituting this identity into the definition of $\mathcal{T}$ gives

$$\begin{aligned} \mathcal{T}(cQ) &= (I-\alpha D)(cQ) + \alpha\gamma DP \begin{bmatrix} \max_a (cQ(1,a)) \\ \vdots \\ \max_a (cQ(|\mathcal{S}|,a)) \end{bmatrix} \\ &= c(I-\alpha D)Q + c\alpha\gamma DP \begin{bmatrix} \max_a Q(1,a) \\ \vdots \\ \max_a Q(|\mathcal{S}|,a) \end{bmatrix} \\ &= c\mathcal{T}(Q). \end{aligned}$$

□

The following three lemmas separate the entrywise-positive case, the weighted-norm estimate, and the extension to a general transition kernel.

Lemma 21. *Suppose that Assumption 1 holds and that P is entrywise positive. Then there exist $\bar{Q} \gg 0$, $\lambda > 0$, and a deterministic policy $\pi_{\bar{Q}} \in \Theta$ such that*

$$M_{\pi_{\bar{Q}}} \bar{Q} = \lambda \bar{Q}$$

and

$$M_\mu \bar{Q} \leq \lambda \bar{Q}$$

for every stochastic policy μ.

Proof. The vectors and matrices used below satisfy

$$Q, \bar{Q}, \mathcal{T}(Q), \mathcal{T}(\bar{Q}) \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \quad D, M_\mu, M_{\pi_{\bar{Q}}} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)},$$
$$P \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times|\mathcal{S}|}, \qquad \Pi^\mu, \Pi^{\pi_{\bar{Q}}} \in \mathbb{R}^{|\mathcal{S}|\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

The map $\mathcal{T}$ is continuous because it is formed from linear maps and finitely many coordinatewise maxima. The two preceding lemmas establish its monotonicity and positive homogeneity.

For every $Q \in \Delta_{|\mathcal{S}|\,|\mathcal{A}|}$,

$$\sum_{s\in\mathcal{S}} \max_{a\in\mathcal{A}} Q(s,a) \geq \frac{1}{|\mathcal{A}|} \sum_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} Q(s,a)$$
$$= \frac{1}{|\mathcal{A}|} > 0.$$

Thus the vector of state-wise maxima is nonzero and entrywise nonnegative. Since P is entrywise positive, for every $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\left[P \begin{bmatrix} \max_{u\in\mathcal{A}} Q(1,u) \\ \vdots \\ \max_{u\in\mathcal{A}} Q(|\mathcal{S}|,u) \end{bmatrix}\right]_{(s,a)}$$
$$= \sum_{s'\in\mathcal{S}} P(s' \mid s,a) \max_{u\in\mathcal{A}} Q(s',u) > 0.$$

Moreover, $1 - \alpha d(s,a) \geq 0$, $d(s,a) > 0$, and $\alpha\gamma > 0$. Hence

$$\mathcal{T}(Q)(s,a) = (1-\alpha d(s,a))Q(s,a)$$
$$+ \alpha\gamma d(s,a) \sum_{s'\in\mathcal{S}} P(s' \mid s,a) \max_{u\in\mathcal{A}} Q(s',u) > 0.$$

Therefore every component of $\mathcal{T}(Q)$ is strictly positive, so $\mathbf{1}_{|\mathcal{S}|\,|\mathcal{A}|}^\top \mathcal{T}(Q) > 0$.

Consider the normalized map

$$Q \longmapsto \frac{\mathcal{T}(Q)}{\mathbf{1}_{|\mathcal{S}|\,|\mathcal{A}|}^\top \mathcal{T}(Q)}.$$

Its numerator and denominator are continuous, the denominator is strictly positive, and its value satisfies

$$\frac{\mathcal{T}(Q)}{\mathbf{1}_{|\mathcal{S}|\,|\mathcal{A}|}^\top \mathcal{T}(Q)} \gg 0, \qquad \mathbf{1}_{|\mathcal{S}|\,|\mathcal{A}|}^\top \frac{\mathcal{T}(Q)}{\mathbf{1}_{|\mathcal{S}|\,|\mathcal{A}|}^\top \mathcal{T}(Q)} = 1.$$

Thus this map is a continuous self-map of $\Delta_{|\mathcal{S}|\,|\mathcal{A}|}$, with image in the simplex interior. By Lemma 20, there exists $\bar{Q} \in \Delta_{|\mathcal{S}|\,|\mathcal{A}|}$ such that

$$\bar{Q} = \frac{\mathcal{T}(\bar{Q})}{\mathbf{1}_{|\mathcal{S}|\,|\mathcal{A}|}^\top \mathcal{T}(\bar{Q})}. \tag{47}$$

Since the image lies in the interior, $\bar{Q} \gg 0$. Define

$$\lambda := \mathbf{1}_{|\mathcal{S}|\,|\mathcal{A}|}^\top \mathcal{T}(\bar{Q}) > 0.$$

The fixed-point identity in Equation (47) gives

$$\mathcal{T}(\bar{Q}) = \lambda \bar{Q}.$$

Choose a deterministic policy

$$\pi_{\bar{Q}}(s) \in \arg\max_{a \in \mathcal{A}} \bar{Q}(s,a), \qquad s \in \mathcal{S}.$$

Then

$$(\Pi^{\pi_{\bar{Q}}} \bar{Q})(s) = \max_{a \in \mathcal{A}} \bar{Q}(s,a).$$

For every stochastic policy μ,

$$\begin{aligned}(\Pi^{\mu} \bar{Q})(s) &= \sum_{a \in \mathcal{A}} \mu(a \mid s) \bar{Q}(s,a) \\ &\le \sum_{a \in \mathcal{A}} \mu(a \mid s) \max_{u \in \mathcal{A}} \bar{Q}(s,u) \\ &= (\Pi^{\pi_{\bar{Q}}} \bar{Q})(s).\end{aligned}$$

Because DP is nonnegative,

$$\begin{aligned}M_{\pi_{\bar{Q}}} \bar{Q} &= (I - \alpha D)\bar{Q} + \alpha\gamma DP \Pi^{\pi_{\bar{Q}}} \bar{Q} \\ &= \mathcal{T}(\bar{Q}) = \lambda \bar{Q},\end{aligned}$$

and

$$\begin{aligned}M_{\mu} \bar{Q} &= (I - \alpha D)\bar{Q} + \alpha\gamma DP \Pi^{\mu} \bar{Q} \\ &\le (I - \alpha D)\bar{Q} + \alpha\gamma DP \Pi^{\pi_{\bar{Q}}} \bar{Q} \\ &= \lambda \bar{Q}.\end{aligned}$$

□

Lemma 22. *Under the conditions of Lemma 21,*

$$\rho(\mathcal{M}_\alpha) = \lambda = \rho(M_{\pi_{\bar{Q}}}) = \max_{\pi \in \Theta} \rho(M_\pi).$$

Proof. The vectors and modes used below satisfy

$$x, z, |x|, |z|, \bar{Q} \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \qquad M_\pi \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|) \times (|\mathcal{S}|\,|\mathcal{A}|)}.$$

Define the weighted maximum norm

$$\|x\|_{\infty,\bar{Q}} := \max_{(s,a) \in \mathcal{S} \times \mathcal{A}} \frac{|x(s,a)|}{\bar{Q}(s,a)}.$$

For a vector z, the notation $|z|$ denotes the vector of componentwise absolute values. Every mode M_π is entrywise nonnegative. Therefore, for each $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}|(M_\pi x)(s,a)| &= \left| \sum_{(\bar{s},\bar{a}) \in \mathcal{S} \times \mathcal{A}} [M_\pi]_{(s,a),(\bar{s},\bar{a})} x(\bar{s},\bar{a}) \right| \\ &\le \sum_{(\bar{s},\bar{a}) \in \mathcal{S} \times \mathcal{A}} [M_\pi]_{(s,a),(\bar{s},\bar{a})} |x(\bar{s},\bar{a})| \\ &= (M_\pi |x|)(s,a).\end{aligned}$$

Moreover, the definition of the weighted norm gives, for every $(\bar{s},\bar{a}) \in \mathcal{S} \times \mathcal{A}$,

$$\frac{|x(\bar{s},\bar{a})|}{\bar{Q}(\bar{s},\bar{a})} \le \|x\|_{\infty,\bar{Q}}, \qquad |x(\bar{s},\bar{a})| \le \|x\|_{\infty,\bar{Q}} \bar{Q}(\bar{s},\bar{a}).$$

Using the nonnegativity of M_π and Lemma 21, we obtain, for each $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\begin{aligned}|(M_\pi x)(s,a)| &\le \sum_{(\bar{s},\bar{a})\in\mathcal{S}\times\mathcal{A}} [M_\pi]_{(s,a),(\bar{s},\bar{a})}|x(\bar{s},\bar{a})| \\ &\le \|x\|_{\infty,\bar{Q}} \sum_{(\bar{s},\bar{a})\in\mathcal{S}\times\mathcal{A}} [M_\pi]_{(s,a),(\bar{s},\bar{a})}\bar{Q}(\bar{s},\bar{a}) \\ &= \|x\|_{\infty,\bar{Q}}(M_\pi\bar{Q})(s,a) \\ &\le \lambda\|x\|_{\infty,\bar{Q}}\bar{Q}(s,a).\end{aligned}$$

Equivalently, using Lemma 21 leads to

$$|M_\pi x| \le M_\pi|x| \le \|x\|_{\infty,\bar{Q}} M_\pi\bar{Q} \le \lambda\|x\|_{\infty,\bar{Q}}\bar{Q}.$$

Since $\bar{Q}(s,a) > 0$, division by $\bar{Q}(s,a)$ gives

$$\frac{|(M_\pi x)(s,a)|}{\bar{Q}(s,a)} \le \lambda\|x\|_{\infty,\bar{Q}} \qquad \text{for every } (s,a) \in \mathcal{S}\times\mathcal{A}.$$

Taking the maximum over (s,a) yields

$$\begin{aligned}\|M_\pi x\|_{\infty,\bar{Q}} &= \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{|(M_\pi x)(s,a)|}{\bar{Q}(s,a)} \\ &\le \lambda\|x\|_{\infty,\bar{Q}}.\end{aligned}$$

Therefore, by the definition of the induced norm,

$$\begin{aligned}\|M_\pi\|_{\infty,\bar{Q}} &:= \sup_{x\neq 0} \frac{\|M_\pi x\|_{\infty,\bar{Q}}}{\|x\|_{\infty,\bar{Q}}} \\ &\le \lambda.\end{aligned}$$

For any sequence $\pi_1,\dots,\pi_k$, submultiplicativity gives

$$\begin{aligned}\|M_{\pi_k}\cdots M_{\pi_1}\|_{\infty,\bar{Q}} &\le \prod_{r=1}^{k} \|M_{\pi_r}\|_{\infty,\bar{Q}} \\ &\le \lambda^k.\end{aligned}$$

Consequently, the definition of the JSR implies

$$\begin{aligned}\rho(\mathcal{M}_\alpha) &= \lim_{k\to\infty} \max_{\pi_1,\dots,\pi_k\in\Theta} \|M_{\pi_k}\cdots M_{\pi_1}\|_{\infty,\bar{Q}}^{1/k} \\ &\le \lim_{k\to\infty} (\lambda^k)^{1/k} = \lambda.\end{aligned}$$

For each fixed π, the spectral radius is bounded by every induced matrix norm. More explicitly,

$$\begin{aligned}\rho(M_\pi) &= \lim_{k\to\infty} \|M_\pi^k\|_{\infty,\bar{Q}}^{1/k} \\ &\le \lim_{k\to\infty} \left(\|M_\pi\|_{\infty,\bar{Q}}^k\right)^{1/k} \\ &= \|M_\pi\|_{\infty,\bar{Q}} \le \lambda, \qquad \pi\in\Theta.\end{aligned}$$

On the other hand, $M_{\pi_{\bar{Q}}}\bar{Q} = \lambda\bar{Q}$, so $\lambda > 0$ is an eigenvalue of $M_{\pi_{\bar{Q}}}$. Hence

$$\rho(M_{\pi_{\bar{Q}}}) \ge |\lambda| = \lambda.$$

Repeating this fixed mode is an admissible switching sequence, and therefore

$$\begin{aligned}\rho(\mathcal{M}_\alpha) &\ge \lim_{k\to\infty} \|M_{\pi_{\bar{Q}}}^k\|_{\infty,\bar{Q}}^{1/k} \\ &= \rho(M_{\pi_{\bar{Q}}}) \\ &\ge \lambda.\end{aligned}$$

Combining the upper and lower bounds gives the complete chain

$$\lambda \le \rho(M_{\pi_Q}) \le \max_{\pi\in\Theta} \rho(M_\pi) \le \rho(\mathcal{M}_\alpha) \le \lambda,$$

which proves

$$\rho(\mathcal{M}_\alpha) = \lambda = \rho(M_{\pi_Q}) = \max_{\pi\in\Theta} \rho(M_\pi).$$

□

Lemma 23. *Under Assumption 1, for a general transition kernel P,*

$$\rho(\mathcal{M}_\alpha) \le \max_{\pi\in\Theta} \rho(M_\pi).$$

Proof. The perturbed kernels, modes, and common vector satisfy

$$P, P_\eta \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times|\mathcal{S}|}, \quad M_\pi, M_\pi^\eta \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}, \quad Q_\eta \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}.$$

For $\eta > 0$, set

$$P_\eta := P + \eta \mathbf{1}_{|\mathcal{S}|\,|\mathcal{A}|} \mathbf{1}_{|\mathcal{S}|}^\top.$$

This auxiliary matrix is entrywise positive and need not be row stochastic. For each $\pi \in \Theta$, define

$$\begin{aligned} M_\pi^\eta &:= I - \alpha D + \alpha\gamma D P_\eta \Pi^\pi \\ &= M_\pi + \alpha\gamma\eta D \mathbf{1}_{|\mathcal{S}|\,|\mathcal{A}|} \mathbf{1}_{|\mathcal{S}|}^\top \Pi^\pi. \end{aligned}$$

The added matrix is nonnegative, so

$$0 \le M_\pi \le M_\pi^\eta \qquad \text{entrywise.}$$

The arguments in Lemmas 21 and 22 use only the entrywise positivity of the matrix replacing P and the nonnegativity of the resulting modes. They do not use row stochasticity. Consequently, there exist $Q_\eta \gg 0$ and

$$\lambda_\eta := \max_{\pi\in\Theta} \rho(M_\pi^\eta)$$

such that

$$M_\pi^\eta Q_\eta \le \lambda_\eta Q_\eta, \qquad \pi \in \Theta.$$

Therefore,

$$M_\pi Q_\eta \le M_\pi^\eta Q_\eta \le \lambda_\eta Q_\eta, \qquad \pi \in \Theta.$$

Repeating the weighted-norm calculation with weight Q_η gives

$$\rho(\mathcal{M}_\alpha) \le \lambda_\eta.$$

Because Θ is finite and the spectral radius is continuous in the matrix entries,

$$\begin{aligned} \lim_{\eta\downarrow 0} \lambda_\eta &= \lim_{\eta\downarrow 0} \max_{\pi\in\Theta} \rho(M_\pi^\eta) \\ &= \max_{\pi\in\Theta} \lim_{\eta\downarrow 0} \rho(M_\pi^\eta) \\ &= \max_{\pi\in\Theta} \rho(M_\pi). \end{aligned}$$

Letting $\eta \downarrow 0$ proves the result. □

The preceding lemmas now yield the structural theorem.

Restatement (Theorem 3)**.** *Under Assumption 1,*

$$\rho(\mathcal{M}_\alpha) = \max_{\pi\in\Theta} \rho(M_\pi).$$

Proof. Every deterministic-policy mode below satisfies

$$M_\pi \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

If P is entrywise positive, Lemmas 21 and 22 prove the result. For a general transition kernel, Lemma 23 gives

$$\rho(\mathcal{M}_\alpha) \le \max_{\pi\in\Theta} \rho(M_\pi).$$

For every fixed $\pi \in \Theta$, the repeated product M_π^k is admissible. Therefore,

$$\rho(\mathcal{M}_\alpha) \ge \lim_{k\to\infty} \|M_\pi^k\|^{1/k} = \rho(M_\pi).$$

Taking the maximum over π proves Equation (7). □

K Proof of Theorem 4

We first restate the LP characterization and then prove its two inequalities.

Restatement (Theorem 4). *Under Assumption 1, $\rho(\mathcal{M}_\alpha)$ equals the infimum of all β for which there exist $q \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$ and $v \in \mathbb{R}^{|\mathcal{S}|}$ satisfying*

$$q \ge \mathbf{1}_{|\mathcal{S}|\,|\mathcal{A}|}, \qquad q \le \mathbf{1}_{|\mathcal{A}|} \otimes v, \qquad (I - \alpha D)q + \alpha\gamma DPv \le \beta q,$$

with all inequalities interpreted entrywise.

For this proof, let $\beta_\star$ denote the infimum in the theorem statement. The next lemma supplies the strict-feasibility direction used in the proof of Theorem 4.

Lemma 24. *For every fixed $\beta > \rho(\mathcal{M}_\alpha)$, Equation (8) is feasible.*

Proof. The vectors, kernels, and modes used below satisfy

$$Q_\eta, q \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \quad v \in \mathbb{R}^{|\mathcal{S}|}, \quad P, P_\eta \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times|\mathcal{S}|},$$
$$M_\pi, M_\pi^\eta \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

Fix $\beta > \rho(\mathcal{M}_\alpha)$. In the perturbation argument used in the proof of Theorem 3,

$$\lambda_\eta = \max_{\pi\in\Theta} \rho(M_\pi^\eta) \longrightarrow \rho(\mathcal{M}_\alpha) \qquad \text{as } \eta \downarrow 0.$$

Choose $\eta > 0$ sufficiently small that $\lambda_\eta < \beta$, and let $Q_\eta \gg 0$ be the associated common positive vector. Define

$$q := \frac{Q_\eta}{\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} Q_\eta(s,a)}, \qquad v(s) := \max_{a\in\mathcal{A}} q(s,a).$$

For every $(s,a) \in \mathcal{S}\times\mathcal{A}$, the definition of q gives

$$q(s,a) = \frac{Q_\eta(s,a)}{\min_{(\bar{s},\bar{a})\in\mathcal{S}\times\mathcal{A}} Q_\eta(\bar{s},\bar{a})} \ge 1, \tag{48}$$

and the definition of v gives

$$q(s,a) \le \max_{u\in\mathcal{A}} q(s,u) = v(s). \tag{49}$$

Equivalently, Equations (48) and (49) can be written in vector form as

$$q \ge \mathbf{1}_{|\mathcal{S}|\,|\mathcal{A}|}, \qquad q \le \mathbf{1}_{|\mathcal{A}|} \otimes v. \tag{50}$$

Choose a deterministic policy

$$\pi_q(s) \in \arg\max_{a\in\mathcal{A}} q(s,a).$$

By construction, $\Pi^{\pi_q} q = v$. Since $P \le P_\eta$ entrywise,

$$\begin{aligned}(I - \alpha D)q + \alpha\gamma DPv &= M_{\pi_q} q \\ &\le M^{\eta}_{\pi_q} q \\ &\le \lambda_\eta q \\ &< \beta q.\end{aligned}$$

Thus all three inequalities in Equation (8) hold, proving feasibility. □

We now complete the proof of Theorem 4.

Proof. The vectors and matrices used below satisfy

$$q, x \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \quad v, \Pi^\mu q \in \mathbb{R}^{|\mathcal{S}|}, \quad A, M_\mu, \widetilde{M}, M_i \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

Suppose first that (β, q, v) satisfies Equation (8). From Equation (8a), $q \gg 0$. For any stochastic policy μ, Equation (8b) gives, for every $s \in \mathcal{S}$,

$$\begin{aligned}(\Pi^\mu q)(s) &= \sum_{a\in\mathcal{A}} \mu(a \mid s) q(s,a) \\ &\le \sum_{a\in\mathcal{A}} \mu(a \mid s) v(s) \\ &= v(s).\end{aligned}$$

Thus $\Pi^\mu q \le v$, and nonnegativity of DP, together with Equation (8c), yields

$$\begin{aligned}M_\mu q &= (I - \alpha D)q + \alpha\gamma DP\Pi^\mu q \\ &\le (I - \alpha D)q + \alpha\gamma DPv \\ &\le \beta q.\end{aligned}$$

Define

$$\|x\|_{\infty,q} := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \frac{|x(s,a)|}{q(s,a)}. \tag{51}$$

The induced matrix norm associated with Equation (51) is

$$\|A\|_{\infty,q} := \sup_{x\neq 0} \frac{\|Ax\|_{\infty,q}}{\|x\|_{\infty,q}}. \tag{52}$$

Since every M_μ is nonnegative, for each $(s,a) \in \mathcal{S}\times\mathcal{A}$,

$$\begin{aligned}|(M_\mu x)(s,a)| &= \left| \sum_{(\bar{s},\bar{a})\in\mathcal{S}\times\mathcal{A}} [M_\mu]_{(s,a),(\bar{s},\bar{a})} x(\bar{s},\bar{a}) \right| \\ &\le \sum_{(\bar{s},\bar{a})\in\mathcal{S}\times\mathcal{A}} [M_\mu]_{(s,a),(\bar{s},\bar{a})} |x(\bar{s},\bar{a})| \\ &= (M_\mu |x|)(s,a).\end{aligned} \tag{53}$$

The definition in Equation (51) gives, for every $(\bar{s},\bar{a}) \in \mathcal{S}\times\mathcal{A}$,

$$\frac{|x(\bar{s},\bar{a})|}{q(\bar{s},\bar{a})} \le \|x\|_{\infty,q}, \qquad |x(\bar{s},\bar{a})| \le \|x\|_{\infty,q} q(\bar{s},\bar{a}). \tag{54}$$

Therefore, using Equations (53) and (54) and then $M_\mu q \le \beta q$, we get

$$\begin{aligned}|(M_\mu x)(s,a)| &\le (M_\mu |x|)(s,a) \\ &= \sum_{(\bar{s},\bar{a})\in\mathcal{S}\times\mathcal{A}} [M_\mu]_{(s,a),(\bar{s},\bar{a})} |x(\bar{s},\bar{a})| \\ &\le \|x\|_{\infty,q} \sum_{(\bar{s},\bar{a})\in\mathcal{S}\times\mathcal{A}} [M_\mu]_{(s,a),(\bar{s},\bar{a})} q(\bar{s},\bar{a}) \\ &= \|x\|_{\infty,q} (M_\mu q)(s,a) \\ &\le \beta \|x\|_{\infty,q} q(s,a).\end{aligned} \tag{55}$$

Since $q(s,a) > 0$, division by $q(s,a)$ in Equation (55) yields

$$\frac{|(M_\mu x)(s,a)|}{q(s,a)} \le \beta \|x\|_{\infty,q} \qquad \text{for every } (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{56}$$

Taking the maximum over all state-action pairs in Equation (56) gives

$$\|M_\mu x\|_{\infty,q} \le \beta \|x\|_{\infty,q}. \tag{57}$$

In particular, the same estimate holds for every deterministic-policy mode. For an arbitrary matrix $\widetilde{M} \in \mathrm{co}(\mathcal{M}_\alpha)$, write

$$\widetilde{M} = \sum_{\pi \in \Theta} c_\pi M_\pi, \qquad c_\pi \ge 0, \qquad \sum_{\pi \in \Theta} c_\pi = 1. \tag{58}$$

Then the triangle inequality and Equation (57) give

$$\|\widetilde{M} x\|_{\infty,q} \le \sum_{\pi \in \Theta} c_\pi \|M_\pi x\|_{\infty,q} \le \beta \|x\|_{\infty,q}. \tag{59}$$

Iterating Equation (59) over arbitrary products first gives

$$\|M_{k-1} \cdots M_0 x\|_{\infty,q} \le \beta^k \|x\|_{\infty,q}, \qquad M_i \in \mathrm{co}(\mathcal{M}_\alpha). \tag{60}$$

Using the induced-norm definition in Equation (52), Equation (60) is equivalently written as

$$\|M_{k-1} \cdots M_0\|_{\infty,q} \le \beta^k, \qquad M_i \in \mathrm{co}(\mathcal{M}_\alpha). \tag{61}$$

Hence $\rho(\mathcal{M}_\alpha) \le \beta$ for every feasible β, and therefore

$$\rho(\mathcal{M}_\alpha) \le \beta_\star. \tag{62}$$

Conversely, Lemma 24 shows that Equation (8) is feasible for every fixed $\beta > \rho(\mathcal{M}_\alpha)$. Thus

$$\beta_\star \le \beta \qquad \text{for every } \beta > \rho(\mathcal{M}_\alpha).$$

Letting $\beta \downarrow \rho(\mathcal{M}_\alpha)$ gives $\beta_\star \le \rho(\mathcal{M}_\alpha)$. Combining the two inequalities proves

$$\beta_\star = \rho(\mathcal{M}_\alpha).$$

□

L Proof of Theorem 5

This appendix proves Theorem 5. We use the single-trajectory recursion in Equation (9) and the mixing-time assumption in Assumption 2. Thus, in this appendix, d is the stationary distribution of the behavior-induced chain, not necessarily the law of $X_k = (s_k, a_k)$ at a finite time. We set $D = \mathrm{diag}(d)$, and $d_{\min}, d_{\max}, \rho_{\mathrm{row}}$ are computed from this stationary state-action distribution. The proof first identifies the stationary averaged recursion, then controls the Markovian reference-system noise, and finally repeats the same pathwise coupled linear decomposition used in Appendix I.

Recall the single-trajectory update in Equation (9). Define the sample update vector

$$\widehat{h}_k(Q_k) := e_{X_k} \left(r_{k+1} + \gamma \max_{u \in \mathcal{A}} Q_k(s_{k+1}, u) - Q_k(X_k) \right),$$

where e_{X_k} is the coordinate vector associated with $X_k = (s_k, a_k)$. Define the stationary averaged update

$$h(Q) := D(F(Q) - Q).$$

Then the single-trajectory recursion can be written exactly as

$$Q_{k+1} = Q_k + \alpha h(Q_k) + \alpha w_k, \qquad w_k := \widehat{h}_k(Q_k) - h(Q_k). \tag{63}$$

The following is the Markovian-observation analogue of Lemma 5. The second-moment statement in Lemma 5 is replaced by a pathwise bound, because the Markovian noise w_k is not generally a martingale difference.

Lemma. *Under Assumptions 1 and 2, the single-trajectory iterates satisfy*

$$\|Q_k\|_\infty \le \max\left\{\|Q_0\|_\infty, \frac{R_{\max}}{1-\gamma}\right\}, \qquad \forall k \ge 0.$$

Moreover, for $W_{\max}$ defined in Equation (4),

$$\|\widehat{h}_k(Q_k)\|_2 \le \sqrt{W_{\max}}, \qquad \|h(Q_k)\|_2 \le \sqrt{W_{\max}}, \qquad \|w_k\|_2 \le 2\sqrt{W_{\max}},$$

and therefore $\mathbb{E}[\|w_k\|_2^2] \le 4W_{\max}$ for every $k \ge 0$.

Proof. The table, update, drift, and noise vectors satisfy

$$Q_k, Q_0, \widehat{h}_k(Q_k), h(Q_k), w_k \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}.$$

The boundedness proof is the same induction as in Lemma 5, but we spell it out for the single-trajectory recursion. Let

$$B_0 := \max\left\{\|Q_0\|_\infty, \frac{R_{\max}}{1-\gamma}\right\}.$$

The claim is true at $k = 0$. Suppose $\|Q_k\|_\infty \le B_0$. The sampled one-step target is

$$Y_k := r_{k+1} + \gamma \max_{u\in\mathcal{A}} Q_k(s_{k+1}, u).$$

Since $|r_{k+1}| \le R_{\max}$ and $\max_{u\in\mathcal{A}} |Q_k(s_{k+1}, u)| \le B_0$,

$$|Y_k| \le R_{\max} + \gamma B_0 \le (1-\gamma)B_0 + \gamma B_0 = B_0.$$

Only the coordinate $X_k = (s_k, a_k)$ is updated, and that coordinate satisfies

$$Q_{k+1}(X_k) = (1-\alpha)Q_k(X_k) + \alpha Y_k.$$

Because $0 < \alpha < 1$, this is a convex combination of two values in $[-B_0, B_0]$. All other coordinates are unchanged, so $\|Q_{k+1}\|_\infty \le B_0$. This proves the uniform bound by induction.

For the sample-update bound, use the just-proved inequality to get

$$\begin{aligned}\left|r_{k+1} + \gamma \max_{u\in\mathcal{A}} Q_k(s_{k+1}, u) - Q_k(X_k)\right| &\le R_{\max} + (1+\gamma)B_0 \\ &= \sqrt{W_{\max}}.\end{aligned}$$

Since $\widehat{h}_k(Q_k)$ has only one nonzero coordinate, $\|\widehat{h}_k(Q_k)\|_2 \le \sqrt{W_{\max}}$. For the averaged update, each coordinate of $F(Q_k) - Q_k$ is bounded in absolute value by the same quantity. Therefore

$$\begin{aligned}\|h(Q_k)\|_2^2 &= \sum_{x\in\mathcal{S}\times\mathcal{A}} d(x)^2\,|F(Q_k)(x) - Q_k(x)|^2 \\ &\le W_{\max} \sum_{x\in\mathcal{S}\times\mathcal{A}} d(x)^2 \le W_{\max},\end{aligned}$$

because d is a probability distribution. Finally,

$$\|w_k\|_2 = \|\widehat{h}_k(Q_k) - h(Q_k)\|_2 \le \|\widehat{h}_k(Q_k)\|_2 + \|h(Q_k)\|_2 \le 2\sqrt{W_{\max}},$$

which also gives $\mathbb{E}[\|w_k\|_2^2] \le 4W_{\max}$. □

The stationary centering identity below explains why $h(Q)$ is the correct averaged drift under the behavior-chain stationary distribution.

Lemma 25. *For every fixed table Q,*

$$\sum_{x\in\mathcal{S}\times\mathcal{A}} d(x)\,\mathbb{E}[\widehat{h}_k(Q) \mid X_k = x] = h(Q).$$

This centering identity is with respect to the stationary distribution d and does not require the actual law of X_k to equal d.

Proof. The table, conditional sample update, and stationary mean drift satisfy

$$Q, \widehat{h}_k(Q), h(Q) \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}.$$

Fix Q and $x = (s, a)$. Conditional on $X_k = x$, the only nonzero coordinate of $\widehat{h}_k(Q)$ is the coordinate (s, a), and its conditional mean is

$$\sum_{s' \in \mathcal{S}} P(s' \mid s, a) \left(r(s, a, s') + \gamma \max_{u \in \mathcal{A}} Q(s', u) - Q(s, a) \right) = F(Q)(s, a) - Q(s, a).$$

Multiplying this coordinate by $d(s, a)$ and summing over (s, a) gives $D(F(Q) - Q) = h(Q)$. No assumption on the initial distribution of X_0 is used. □

The noise w_k in Equation (63) is not, in general, a martingale difference with respect to the trajectory filtration, because the Markovian coordinate X_k is correlated with the past and with Q_k. The averaged drift, however, has exactly the same direct SLS form as in the i.i.d. case. Since $h(Q^*) = 0$, Lemma 6 gives, for each k, an adapted stochastic policy μ_k such that

$$F(Q_k) - F(Q^*) = \gamma P \Pi^{\mu_k} (Q_k - Q^*).$$

Therefore

$$h(Q_k) - h(Q^*) = D(\gamma P \Pi^{\mu_k} - I)(Q_k - Q^*).$$

Combining this identity with Equation (63) yields the exact stationary-distribution averaged direct SLS

$$Q_{k+1} - Q^* = M_{\mu_k}(Q_k - Q^*) + \alpha w_k. \tag{64}$$

This is the same direct matrix family as in the i.i.d. analysis; only the noise process is Markovian.

We next solve the Poisson equation for the centered conditional sample update. The solution is denoted by u_Q; for each $x \in \mathcal{S} \times \mathcal{A}$, $u_Q(x) \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$.

Lemma 26. *Under Assumption 2, for every fixed table Q satisfying*

$$\|Q\|_\infty \le \max\left\{ \|Q_0\|_\infty, \frac{R_{\max}}{1 - \gamma} \right\},$$

the Poisson equation

$$u_Q(x) - \sum_{x' \in \mathcal{S} \times \mathcal{A}} P^b(x, x') u_Q(x') = \mathbb{E}[\widehat{h}_k(Q) \mid X_k = x] - h(Q), \qquad x \in \mathcal{S} \times \mathcal{A},$$

has a solution $u_Q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$. Moreover, the solution can be selected so that, for every $x \in \mathcal{S} \times \mathcal{A}$, the map $Q \mapsto u_Q(x)$ is Borel measurable on the bounded set specified above; in fact, the series construction below is continuous there.

Proof. For every $x \in \mathcal{S} \times \mathcal{A}$, the block vectors used below satisfy

$$Q, g_Q(x), u_Q(x) \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}.$$

For a fixed table Q, put

$$g_Q(x) := \mathbb{E}[\widehat{h}_k(Q) \mid X_k = x] - h(Q), \qquad x \in \mathcal{S} \times \mathcal{A}.$$

Each $g_Q(x)$ is a vector in $\mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$. By Lemma 25,

$$\sum_{x \in \mathcal{S} \times \mathcal{A}} d(x) g_Q(x) = 0. \tag{65}$$

Define

$$u_Q(x) := \sum_{\ell=0}^{\infty} \sum_{y \in \mathcal{S} \times \mathcal{A}} \big((P^b)^\ell(x, y) - d(y) \big) g_Q(y). \tag{66}$$

We first check that the series is absolutely convergent. For every x and ℓ,

$$\begin{aligned}
&\left\| \sum_{y \in \mathcal{S}\times\mathcal{A}} \big((P^b)^\ell(x,y) - d(y)\big) g_Q(y) \right\|_2 \\
&\quad \le \sum_{y \in \mathcal{S}\times\mathcal{A}} \big|(P^b)^\ell(x,y) - d(y)\big| \sup_{z \in \mathcal{S}\times\mathcal{A}} \|g_Q(z)\|_2 \\
&\quad = 2 \left\|(P^b)^\ell(x,\cdot) - d\right\|_{\mathrm{TV}} \sup_{z \in \mathcal{S}\times\mathcal{A}} \|g_Q(z)\|_2 \\
&\quad \le 2^{1-\lfloor \ell/t_{\mathrm{mix}} \rfloor} \sup_{z \in \mathcal{S}\times\mathcal{A}} \|g_Q(z)\|_2.
\end{aligned}$$

Since $t_{\mathrm{mix}} \ge 1$,

$$\sum_{\ell=0}^{\infty} 2^{1-\lfloor \ell/t_{\mathrm{mix}} \rfloor} = 2 \sum_{\ell=0}^{\infty} 2^{-\lfloor \ell/t_{\mathrm{mix}} \rfloor} \le 8 t_{\mathrm{mix}}.$$

Moreover, the bound on Q implies, exactly as in the Markovian-observation version of Lemma 5 above, that

$$\sup_{x \in \mathcal{S}\times\mathcal{A}} \|\mathbb{E}[\widehat{h}_k(Q) \mid X_k = x]\|_2 \le \sqrt{W_{\max}}, \qquad \|h(Q)\|_2 \le \sqrt{W_{\max}}.$$

Thus

$$\sup_{x \in \mathcal{S}\times\mathcal{A}} \|g_Q(x)\|_2 \le 2\sqrt{W_{\max}}. \tag{67}$$

The preceding bounds show that the series in Equation (66) is absolutely convergent and that

$$\sup_{x \in \mathcal{S}\times\mathcal{A}} \|u_Q(x)\|_2 \le 8 t_{\mathrm{mix}} \sup_{x \in \mathcal{S}\times\mathcal{A}} \|g_Q(x)\|_2. \tag{68}$$

Because the state and action spaces are finite, the maps $Q \mapsto \mathbb{E}[\widehat{h}_k(Q) \mid X_k = x]$, $Q \mapsto h(Q)$, and hence $Q \mapsto g_Q(x)$, are continuous; the only nonlinearity is the finite maximum in the Bellman operator. The partial sums in Equation (66) are therefore continuous functions of Q. On the bounded set in the lemma, Equation (67) gives a uniform dominating summable series, so the convergence in Equation (66) is uniform on that set. Thus $Q \mapsto u_Q(x)$ is a uniform limit of continuous functions on the bounded set and is Borel measurable, indeed continuous there.

It remains to verify that the constructed u_Q solves the Poisson equation. Applying P^b to Equation (66) and using $dP^b = d$,

$$\begin{aligned}
\sum_{x' \in \mathcal{S}\times\mathcal{A}} P^b(x,x') u_Q(x') &= \sum_{x' \in \mathcal{S}\times\mathcal{A}} P^b(x,x') \sum_{\ell=0}^{\infty} \sum_{y \in \mathcal{S}\times\mathcal{A}} \big((P^b)^\ell(x',y) - d(y)\big) g_Q(y) \\
&= \sum_{\ell=0}^{\infty} \sum_{y \in \mathcal{S}\times\mathcal{A}} \big((P^b)^{\ell+1}(x,y) - d(y)\big) g_Q(y) \\
&= \sum_{\ell=1}^{\infty} \sum_{y \in \mathcal{S}\times\mathcal{A}} \big((P^b)^{\ell}(x,y) - d(y)\big) g_Q(y).
\end{aligned}$$

Therefore

$$\begin{aligned}
&u_Q(x) - \sum_{x' \in \mathcal{S}\times\mathcal{A}} P^b(x,x') u_Q(x') \\
&\quad = \sum_{y \in \mathcal{S}\times\mathcal{A}} \big(\mathbf{1}\{x = y\} - d(y)\big) g_Q(y) \\
&\quad = g_Q(x) - \sum_{y \in \mathcal{S}\times\mathcal{A}} d(y) g_Q(y) = g_Q(x),
\end{aligned}$$

where the last equality uses Equation (65). This is the desired Poisson equation. □

The next lemma turns the blockwise Poisson estimate into a bound for the stacked solution.

Lemma 27. *Under Assumption 2, for every fixed table Q satisfying*

$$\|Q\|_\infty \le \max\left\{\|Q_0\|_\infty, \frac{R_{\max}}{1-\gamma}\right\},$$

the solution constructed in Lemma 26 satisfies

$$\|u_Q\|_2 \le 32|\mathcal{S}|\,|\mathcal{A}|\,t_{\mathrm{mix}}\sqrt{W_{\max}}.$$

Here the norm is the Euclidean norm after stacking the $|\mathcal{S}|\,|\mathcal{A}|$ vectors $u_Q(x) \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$; equivalently, the stacked object lies in $\mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)^2}$.

Proof. The block and stacked Poisson vectors satisfy

$$u_Q(x) \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \qquad u_Q \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)^2}.$$

The proof of Lemma 26 gives Equation (68). Combining that estimate with Equation (67) yields

$$\sup_{x\in\mathcal{S}\times\mathcal{A}} \|u_Q(x)\|_2 \le 16 t_{\mathrm{mix}}\sqrt{W_{\max}}. \tag{69}$$

Because the stacked vector has $|\mathcal{S}|\,|\mathcal{A}|$ blocks, each in $\mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$,

$$\begin{aligned}\|u_Q\|_2^2 &= \sum_{x\in\mathcal{S}\times\mathcal{A}} \|u_Q(x)\|_2^2 \\ &\le |\mathcal{S}|\,|\mathcal{A}|\left(16 t_{\mathrm{mix}}\sqrt{W_{\max}}\right)^2.\end{aligned}$$

Hence

$$\|u_Q\|_2 \le 16\sqrt{|\mathcal{S}|\,|\mathcal{A}|}\,t_{\mathrm{mix}}\sqrt{W_{\max}} \le 32|\mathcal{S}|\,|\mathcal{A}|\,t_{\mathrm{mix}}\sqrt{W_{\max}}.$$

□

The following lemma shows that the Poisson solution changes Lipschitz continuously with the Q-table.

Lemma 28. *Under Assumption 2, for any two tables Q, Q' satisfying*

$$\|Q\|_\infty, \|Q'\|_\infty \le \max\left\{\|Q_0\|_\infty, \frac{R_{\max}}{1-\gamma}\right\},$$

the solutions constructed in Lemma 26 satisfy

$$\|u_Q - u_{Q'}\|_2 \le 32|\mathcal{S}|\,|\mathcal{A}|\,t_{\mathrm{mix}}\|Q - Q'\|_\infty.$$

Proof. The table, block, and stacked vectors used below satisfy

$$Q, Q' \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \quad g_Q(x), g_{Q'}(x), u_Q(x), u_{Q'}(x) \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|},$$

$$u_Q, u_{Q'} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)^2}.$$

For a fixed table Q, write

$$g_Q(x) := \mathbb{E}[\widehat{h}_k(Q) \mid X_k = x] - h(Q), \qquad x \in \mathcal{S}\times\mathcal{A}.$$

For every $x = (s, a)$, the conditional sample update changes by at most

$$\begin{aligned}
&\Big\|\mathbb{E}[\widehat{h}_k(Q) \mid X_k = x] - \mathbb{E}[\widehat{h}_k(Q') \mid X_k = x]\Big\|_2 \\
&\quad \le \sum_{s'} P(s' \mid s, a) \left(\gamma \left|\max_{u \in \mathcal{A}} Q(s', u) - \max_{u \in \mathcal{A}} Q'(s', u)\right| + |Q(s, a) - Q'(s, a)|\right) \\
&\quad \le (1 + \gamma)\|Q - Q'\|_\infty \le 2\|Q - Q'\|_\infty.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\|h(Q) - h(Q')\|_2^2 &= \sum_{x \in \mathcal{S} \times \mathcal{A}} d(x)^2 \left|F(Q)(x) - Q(x) - F(Q')(x) + Q'(x)\right|^2 \\
&\le 4\|Q - Q'\|_\infty^2 \sum_{x \in \mathcal{S} \times \mathcal{A}} d(x)^2 \le 4\|Q - Q'\|_\infty^2.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\sup_{x \in \mathcal{S} \times \mathcal{A}} \|g_Q(x) - g_{Q'}(x)\|_2 &\le 2\|Q - Q'\|_\infty + 2\|Q - Q'\|_\infty \\
&= 4\|Q - Q'\|_\infty.
\end{aligned}$$

Subtracting the two series definitions in Equation (66) gives

$$u_Q(x) - u_{Q'}(x) = \sum_{\ell=0}^{\infty} \sum_{y \in \mathcal{S} \times \mathcal{A}} \big((P^b)^\ell(x, y) - d(y)\big)\big(g_Q(y) - g_{Q'}(y)\big).$$

Applying the same estimate as Equation (68) to $g_Q - g_{Q'}$ gives the blockwise Lipschitz estimate

$$\sup_{x \in \mathcal{S} \times \mathcal{A}} \|u_Q(x) - u_{Q'}(x)\|_2 \le 32 t_{\mathrm{mix}} \|Q - Q'\|_\infty. \tag{70}$$

Stacking over $|\mathcal{S}|\,|\mathcal{A}|$ values of x,

$$\begin{aligned}
\|u_Q - u_{Q'}\|_2 &\le 32\sqrt{|\mathcal{S}|\,|\mathcal{A}|}\, t_{\mathrm{mix}} \|Q - Q'\|_\infty \\
&\le 32|\mathcal{S}|\,|\mathcal{A}|\, t_{\mathrm{mix}} \|Q - Q'\|_\infty.
\end{aligned}$$

□

The next estimates control the fixed reference system driven by the Markovian noise in Equation (64). They separate the proof into three elementary pieces: the fresh transition noise, the martingale part coming from the Poisson equation, and the remaining telescoping term. For a fixed stochastic policy $\bar{\mu}$, we use throughout

$$\|M_{\bar{\mu}}^j\|_\infty \le \rho_{\mathrm{row}}^j, \qquad \|M_{\bar{\mu}}^j\|_2 \le |\mathcal{S}|\,|\mathcal{A}|\,\rho_{\mathrm{row}}^j, \qquad j \ge 0, \tag{71}$$

where the second inequality follows from $\|B\|_2 \le |\mathcal{S}|\,|\mathcal{A}| \|B\|_\infty$ for matrices of dimension $|\mathcal{S}|\,|\mathcal{A}|$.

Lemma 29. *Fix an arbitrary stochastic policy $\bar{\mu}$. Define*

$$\mathcal{F}_0 := \sigma(Q_0, X_0), \qquad \mathcal{F}_i := \sigma(Q_0, X_0, r_1, X_1, \ldots, r_i, X_i), \qquad i \ge 1,$$

and

$$\bar{h}_Q(x) := \mathbb{E}[\widehat{h}_i(Q) \mid X_i = x], \qquad \eta_i := \widehat{h}_i(Q_i) - \bar{h}_{Q_i}(X_i).$$

Under Assumption 2, for every $k \ge 1$,

$$\mathbb{E}\left[\left\|\alpha \sum_{i=0}^{k-1} M_{\bar{\mu}}^{k-1-i} \eta_i\right\|_\infty\right] \le 2|\mathcal{S}|\,|\mathcal{A}| \sqrt{W_{\max}} \sqrt{\frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1 - \gamma)}}. \tag{72}$$

Proof. The martingale increments, filtered sums, and reference modes satisfy

$$\eta_i \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \qquad M_{\bar{\mu}} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

The random variables Q_i and X_i are $\mathcal{F}_i$-measurable, and the next transition is sampled only after time i. Hence

$$\mathbb{E}[\eta_i \mid \mathcal{F}_i] = 0.$$

The pathwise bounds proved at the beginning of this appendix give

$$\|\eta_i\|_2 \le \|\widehat{h}_i(Q_i)\|_2 + \|\bar{h}_{Q_i}(X_i)\|_2 \le 2\sqrt{W_{\max}}.$$

For $0 \le i < j \le k-1$, the martingale-difference property implies

$$\begin{aligned}&\mathbb{E}\Big[\eta_i^\top \left(M_{\bar{\mu}}^{k-1-i}\right)^\top M_{\bar{\mu}}^{k-1-j}\eta_j\Big] \\ &\quad = \mathbb{E}\Big[\eta_i^\top \left(M_{\bar{\mu}}^{k-1-i}\right)^\top M_{\bar{\mu}}^{k-1-j}\mathbb{E}[\eta_j \mid \mathcal{F}_j]\Big] = 0.\end{aligned}$$

Therefore, using martingale orthogonality, $\|z\|_\infty \le \|z\|_2$, and Equation (71),

$$\begin{aligned}&\mathbb{E}\left[\left\|\alpha\sum_{i=0}^{k-1} M_{\bar{\mu}}^{k-1-i}\eta_i\right\|_\infty\right] \\ &\quad\le \left(\alpha^2\sum_{i=0}^{k-1}\|M_{\bar{\mu}}^{k-1-i}\|_2^2\,\mathbb{E}[\|\eta_i\|_2^2]\right)^{1/2} \\ &\quad\le 2|\mathcal{S}|\,|\mathcal{A}|\sqrt{W_{\max}}\left(\alpha^2\sum_{i=0}^{k-1}\rho_{\text{row}}^{2(k-1-i)}\right)^{1/2} \\ &\quad\le 2|\mathcal{S}|\,|\mathcal{A}|\sqrt{W_{\max}}\sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} \\ &\quad\le 2|\mathcal{S}|\,|\mathcal{A}|\sqrt{W_{\max}}\sqrt{\frac{\alpha t_{\text{mix}}}{d_{\min}(1-\gamma)}}.\end{aligned}$$

Here we used $t_{\text{mix}} \ge 1$, $1-\rho_{\text{row}} = \alpha d_{\min}(1-\gamma)$, and $1-\rho_{\text{row}}^2 \ge 1-\rho_{\text{row}}$. □

The next lemma controls the martingale term generated by the Poisson equation.

Lemma 30. *Fix an arbitrary stochastic policy $\bar{\mu}$. For the solution u_Q in Lemma 26, define*

$$h_{i+1} := u_{Q_i}(X_{i+1}) - \sum_{x\in\mathcal{S}\times\mathcal{A}} P^b(X_i, x)u_{Q_i}(x).$$

Under Assumption 2, for every $k \ge 1$,

$$\mathbb{E}\left[\left\|\alpha\sum_{i=0}^{k-1} M_{\bar{\mu}}^{k-1-i}h_{i+1}\right\|_\infty\right] \le 64|\mathcal{S}|^2|\mathcal{A}|^2\sqrt{W_{\max}}\,t_{\text{mix}}\sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}}. \tag{73}$$

Proof. The Poisson blocks, martingale increments, and reference modes satisfy

$$u_Q(x), h_{i+1} \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \qquad M_{\bar{\mu}} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

With the filtration $\mathcal{F}_i$ from Lemma 29, the random variable h_{i+1} is $\mathcal{F}_{i+1}$-measurable and satisfies

$$\mathbb{E}[h_{i+1} \mid \mathcal{F}_i] = 0.$$

The table Q_i is adaptive and can be correlated with X_i. Therefore we do not freeze Q_i inside a stationary covariance estimate. Instead, we use the pathwise Poisson bound. By Lemma 27, for every admissible table Q,

$$\sup_{x\in\mathcal{S}\times\mathcal{A}} \|u_Q(x)\|_2 \le \|u_Q\|_2 \le 32|\mathcal{S}|\,|\mathcal{A}|\,t_{\mathrm{mix}}\sqrt{W_{\max}}.$$

Consequently, almost surely,

$$\begin{aligned}\|h_{i+1}\|_2 &\le \|u_{Q_i}(X_{i+1})\|_2 + \left\|\sum_{x\in\mathcal{S}\times\mathcal{A}} P^b(X_i,x)u_{Q_i}(x)\right\|_2 \\ &\le 64|\mathcal{S}|\,|\mathcal{A}|\,t_{\mathrm{mix}}\sqrt{W_{\max}}.\end{aligned}$$

Martingale orthogonality and Equation (71) then give

$$\begin{aligned}&\mathbb{E}\left[\left\|\alpha\sum_{i=0}^{k-1} M_{\bar\mu}^{k-1-i}h_{i+1}\right\|_\infty\right] \\ &\quad\le \left(\alpha^2\sum_{i=0}^{k-1}\|M_{\bar\mu}^{k-1-i}\|_2^2\,4096|\mathcal{S}|^2|\mathcal{A}|^2t_{\mathrm{mix}}^2W_{\max}\right)^{1/2} \\ &\quad\le 64|\mathcal{S}|^2|\mathcal{A}|^2\sqrt{W_{\max}}\,t_{\mathrm{mix}}\left(\alpha^2\sum_{i=0}^{k-1}\rho_{\mathrm{row}}^{2(k-1-i)}\right)^{1/2} \\ &\quad\le 64|\mathcal{S}|^2|\mathcal{A}|^2\sqrt{W_{\max}}\,t_{\mathrm{mix}}\sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}}.\end{aligned}$$

Here we used $1-\rho_{\mathrm{row}} = \alpha d_{\min}(1-\gamma)$ and $1-\rho_{\mathrm{row}}^2 \ge 1-\rho_{\mathrm{row}}$. □

The final component bounds the telescoping remainder caused by changes in Q_i.

Lemma 31. *Fix an arbitrary stochastic policy $\bar\mu$, and define*

$$T_k := \alpha\sum_{i=0}^{k-1} M_{\bar\mu}^{k-1-i}\big(u_{Q_i}(X_i) - u_{Q_i}(X_{i+1})\big).$$

Under Assumption 2, for every $k \ge 1$,

$$\mathbb{E}[\|T_k\|_\infty] \le 64|\mathcal{S}|\,|\mathcal{A}|\sqrt{W_{\max}}\frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)}. \tag{74}$$

Proof. The telescoping term, Poisson blocks, and reference modes satisfy

$$T_k, u_{Q_i}(X_i) \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \qquad M_{\bar\mu} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|)\times(|\mathcal{S}|\,|\mathcal{A}|)}.$$

Put $n := |\mathcal{S}|\,|\mathcal{A}|$, $\delta := d_{\min}(1-\gamma)$, and $\rho := \rho_{\mathrm{row}} = 1-\alpha\delta$. If $n = 1$, then $(P^b)^\ell(x,\cdot) = d$ for the unique state-action pair x and every $\ell \ge 0$. Hence the series in Equation (66) gives $u_Q \equiv 0$, so $T_k = 0$ and the claim is immediate. We may therefore assume $n \ge 2$. Since d is a probability distribution on at least two points,

$$d_{\max} + \delta \le d_{\max} + d_{\min} \le 1. \tag{75}$$

A direct summation by parts gives

$$\begin{aligned}T_k = {}& \alpha M_{\bar\mu}^{k-1}u_{Q_0}(X_0) - \alpha u_{Q_{k-1}}(X_k) \\ &+ \alpha\sum_{i=1}^{k-1}\left(M_{\bar\mu}^{k-1-i}u_{Q_i}(X_i) - M_{\bar\mu}^{k-i}u_{Q_{i-1}}(X_i)\right).\end{aligned} \tag{76}$$

For each summand in the last line,

$$M_{\bar{\mu}}^{k-1-i} u_{Q_i}(X_i) - M_{\bar{\mu}}^{k-i} u_{Q_{i-1}}(X_i)$$
$$= M_{\bar{\mu}}^{k-1-i}(I - M_{\bar{\mu}}) u_{Q_i}(X_i) + M_{\bar{\mu}}^{k-i}\big(u_{Q_i} - u_{Q_{i-1}}\big)(X_i).$$

We now use the blockwise Poisson estimates rather than the coarser stacked bounds. From Equation (69),

$$\|u_{Q_i}(X_i)\|_\infty \le 16 t_{\mathrm{mix}} \sqrt{W_{\max}} =: U.$$

Also,

$$I - M_{\bar{\mu}} = \alpha D(I - \gamma P \Pi^{\bar{\mu}}), \qquad \|I - M_{\bar{\mu}}\|_\infty \le 2\alpha d_{\max},$$

and $\|M_{\bar{\mu}}^j\|_\infty \le \rho^j$. Hence the boundary terms and the terms containing $I - M_{\bar{\mu}}$ are bounded pathwise by

$$\alpha U(\rho^{k-1} + 1) + 2\alpha^2 d_{\max} U \sum_{i=1}^{k-1} \rho^{k-1-i}$$
$$\le \alpha U \left(2 + \frac{2 d_{\max}}{\delta}\right) = 32 \alpha t_{\mathrm{mix}} \sqrt{W_{\max}} \frac{d_{\max} + \delta}{\delta} \le 32 \sqrt{W_{\max}} \frac{\alpha t_{\mathrm{mix}}}{\delta},$$

where the last inequality uses Equation (75).

It remains to bound the adaptive-change terms. By Equation (70) and the pathwise sample-update bound,

$$\|\big(u_{Q_i} - u_{Q_{i-1}}\big)(X_i)\|_\infty \le 32 t_{\mathrm{mix}} \|Q_i - Q_{i-1}\|_\infty$$
$$\le 32 \alpha t_{\mathrm{mix}} \sqrt{W_{\max}},$$

because $Q_i = Q_{i-1} + \alpha \widehat{h}_{i-1}(Q_{i-1})$ and $\|\widehat{h}_{i-1}(Q_{i-1})\|_2 \le \sqrt{W_{\max}}$. Therefore

$$\alpha \sum_{i=1}^{k-1} \big\| M_{\bar{\mu}}^{k-i}\big(u_{Q_i} - u_{Q_{i-1}}\big)(X_i) \big\|_\infty$$
$$\le 32 \alpha^2 t_{\mathrm{mix}} \sqrt{W_{\max}} \sum_{i=1}^{k-1} \rho^{k-i} \le 32 \sqrt{W_{\max}} \frac{\alpha t_{\mathrm{mix}}}{\delta}.$$

Combining the two pathwise estimates gives

$$\|T_k\|_\infty \le 64 \sqrt{W_{\max}} \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)} \le 64 |\mathcal{S}|\,|\mathcal{A}| \sqrt{W_{\max}} \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)}.$$

Taking expectations proves Equation (74). □

Combining the preceding three components gives a bound for the fixed-policy reference noise recursion.

Lemma 32. *Fix an arbitrary stochastic policy $\bar{\mu}$. Consider the fixed reference noise recursion*

$$y_{k+1} = M_{\bar{\mu}} y_k + \alpha w_k, \qquad y_0 = 0.$$

Under Assumption 2, for every $k \ge 0$,

$$\mathbb{E}[\|y_k\|_\infty] \le 128 |\mathcal{S}|^2 |\mathcal{A}|^2 \sqrt{W_{\max}} \left(t_{\mathrm{mix}} \sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)} \right). \tag{77}$$

Proof. The reference-noise, Poisson, and martingale vectors and the fixed mode satisfy

$$y_k, w_i, g_Q(x), \eta_i, h_{i+1}, u_Q(x) \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \qquad M_{\bar{\mu}} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|) \times (|\mathcal{S}|\,|\mathcal{A}|)}.$$

The case $k = 0$ is immediate from $y_0 = 0$. Assume $k \geq 1$. Unrolling the fixed reference recursion gives

$$y_k = \alpha \sum_{i=0}^{k-1} M_{\bar{\mu}}^{k-1-i} w_i.$$

Use the filtration $\mathcal{F}_i$ and $\bar{h}_Q$ from Lemma 29. Put

$$g_Q(x) := \bar{h}_Q(x) - h(Q).$$

Then

$$w_i = \underbrace{\widehat{h}_i(Q_i) - \bar{h}_{Q_i}(X_i)}_{\eta_i} + g_{Q_i}(X_i). \tag{78}$$

By Lemma 26,

$$g_{Q_i}(X_i) = u_{Q_i}(X_i) - \sum_{x \in \mathcal{S} \times \mathcal{A}} P^b(X_i, x) u_{Q_i}(x).$$

With h_{i+1} defined in Lemma 30, this gives the one-step decomposition

$$g_{Q_i}(X_i) = u_{Q_i}(X_i) - u_{Q_i}(X_{i+1}) + h_{i+1}. \tag{79}$$

Substituting Equations (78) and (79) into the unrolled formula for y_k gives

$$\begin{aligned} y_k = \alpha \sum_{i=0}^{k-1} M_{\bar{\mu}}^{k-1-i} \eta_i + \alpha \sum_{i=0}^{k-1} M_{\bar{\mu}}^{k-1-i} h_{i+1} \\ + \alpha \sum_{i=0}^{k-1} M_{\bar{\mu}}^{k-1-i} \big(u_{Q_i}(X_i) - u_{Q_i}(X_{i+1})\big). \end{aligned}$$

The three terms are bounded by Lemmas 29 to 31. Adding those bounds, using $t_{\mathrm{mix}} \geq 1$ and $|\mathcal{S}|\,|\mathcal{A}| \geq 1$, and absorbing the numerical constants into the stated factor $128|\mathcal{S}|^2|\mathcal{A}|^2$ proves Equation (77). □

We now prove the Markovian coupled linear bound.

Restatement (Theorem 5)**.** *Let Assumptions 1 and 2 hold, with d equal to the stationary distribution of the behavior-induced chain. Fix $\varepsilon > 0$ such that*

$$\beta_\varepsilon := \rho(\mathcal{M}_\alpha) + \varepsilon < 1,$$

and let q and v satisfy Equation (8) with $\beta = \beta_\varepsilon$. Let K_{β_ε} be the product-growth constant from Lemma 9, and let $W_{\max}$ be defined in Equation (4). Then, for every $k \geq 0$,

$$\begin{aligned} \mathbb{E}[\|Q_k - Q^*\|_\infty] \leq \left(1 + \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)}\right) 128|\mathcal{S}|^2|\mathcal{A}|^2 \sqrt{W_{\max}} \\ \times \left(t_{\mathrm{mix}} \sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)}\right) \\ + K_{\beta_\varepsilon} \beta_\varepsilon^k \|Q_0 - Q^*\|_\infty \\ + 2\alpha\gamma d_{\max} K_{\beta_\varepsilon}^2 k \beta_\varepsilon^{k-1} \|Q_0 - Q^*\|_\infty, \end{aligned}$$

where $k\beta_\varepsilon^{k-1} = 0$ when $k = 0$.

Proof. The reference, residual, and Markovian-noise vectors and active modes satisfy

$$x_k, z_k, y_k, w_k, p_k, u_k, v_k, Q_k - Q^* \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}, \qquad M_{\bar{\mu}}, M_{\mu_k} \in \mathbb{R}^{(|\mathcal{S}|\,|\mathcal{A}|) \times (|\mathcal{S}|\,|\mathcal{A}|)}.$$

The proof follows the coupled linear argument, with Lemma 32 replacing the i.i.d. covariance recursion. The product-growth estimate used below is the definition of K_{β_ε} in Equation (5), with Lemma 9 ensuring $K_{\beta_\varepsilon} \leq \kappa_q$.

Fix an arbitrary stochastic policy $\bar{\mu}$. Introduce the reference system driven by the same Markovian noise as Equation (64):

$$x_{k+1} = M_{\bar{\mu}} x_k + \alpha w_k, \qquad x_0 = Q_0 - Q^*.$$

Decompose

$$x_k = z_k + y_k,$$

where

$$z_{k+1} = M_{\bar{\mu}} z_k, \qquad z_0 = Q_0 - Q^*,$$

and

$$y_{k+1} = M_{\bar{\mu}} y_k + \alpha w_k, \qquad y_0 = 0.$$

Since $M_{\bar{\mu}} \in \mathrm{co}(\mathcal{M}_\alpha)$, the definition of K_{β_ε} gives

$$\|z_k\|_\infty \le K_{\beta_\varepsilon} \beta_\varepsilon^k \|Q_0 - Q^*\|_\infty. \tag{80}$$

By Lemma 32,

$$\mathbb{E}[\|y_k\|_\infty] \le 128 |\mathcal{S}|^2 |\mathcal{A}|^2 \sqrt{W_{\max}} \left(t_{\mathrm{mix}} \sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)} \right). \tag{81}$$

Subtract the reference system from the exact direct SLS Equation (64). With $p_k := Q_k - Q^* - x_k$,

$$\begin{aligned} p_{k+1} &= (Q_{k+1} - Q^*) - x_{k+1} \\ &= M_{\mu_k}(Q_k - Q^*) + \alpha w_k - M_{\bar{\mu}} x_k - \alpha w_k \\ &= M_{\mu_k} p_k + (M_{\mu_k} - M_{\bar{\mu}}) x_k. \end{aligned}$$

The noise cancels pathwise. Moreover,

$$M_{\mu_k} - M_{\bar{\mu}} = \alpha\gamma D P (\Pi^{\mu_k} - \Pi^{\bar{\mu}}),$$

and hence

$$\|M_{\mu_k} - M_{\bar{\mu}}\|_\infty \le 2\alpha\gamma d_{\max}. \tag{82}$$

Because $x_0 = Q_0 - Q^*$, we have $p_0 = 0$.

Split $p_k = u_k + v_k$, where

$$u_{k+1} = M_{\mu_k} u_k + (M_{\mu_k} - M_{\bar{\mu}}) z_k, \qquad u_0 = 0,$$

and

$$v_{k+1} = M_{\mu_k} v_k + (M_{\mu_k} - M_{\bar{\mu}}) y_k, \qquad v_0 = 0.$$

For u_k, unroll the recursion and use the JSR product-growth bound twice:

$$\begin{aligned} \|u_k\|_\infty &\le \sum_{i=0}^{k-1} K_{\beta_\varepsilon} \beta_\varepsilon^{k-1-i} \cdot 2\alpha\gamma d_{\max} \cdot K_{\beta_\varepsilon} \beta_\varepsilon^i \|Q_0 - Q^*\|_\infty \\ &= 2\alpha\gamma d_{\max} K_{\beta_\varepsilon}^2 k \beta_\varepsilon^{k-1} \|Q_0 - Q^*\|_\infty. \end{aligned}$$

Thus

$$\mathbb{E}[\|u_k\|_\infty] \le 2\alpha\gamma d_{\max} K_{\beta_\varepsilon}^2 k \beta_\varepsilon^{k-1} \|Q_0 - Q^*\|_\infty. \tag{83}$$

For v_k, use row contraction rather than the JSR product bound. Since $\|M_\mu\|_\infty \le \rho_{\mathrm{row}}$ for every stochastic policy μ,

$$\begin{aligned} \mathbb{E}[\|v_k\|_\infty] &\le \sum_{i=0}^{k-1} \rho_{\mathrm{row}}^{k-1-i} \cdot 2\alpha\gamma d_{\max} \cdot \mathbb{E}[\|y_i\|_\infty] \\ &\le 2\alpha\gamma d_{\max} 128 |\mathcal{S}|^2 |\mathcal{A}|^2 \sqrt{W_{\max}} \left(t_{\mathrm{mix}} \sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)} \right) \sum_{i=0}^{k-1} \rho_{\mathrm{row}}^{k-1-i} \\ &\le \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)} 128 |\mathcal{S}|^2 |\mathcal{A}|^2 \sqrt{W_{\max}} \left(t_{\mathrm{mix}} \sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)} \right), \end{aligned}$$

where the last step uses $1-\rho_{\mathrm{row}} = \alpha d_{\min}(1-\gamma)$. Therefore

$$\mathbb{E}[\|v_k\|_\infty] \le \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)} 128|\mathcal{S}|^2|\mathcal{A}|^2\sqrt{W_{\max}}\left(t_{\mathrm{mix}}\sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)}\right). \tag{84}$$

Finally,

$$Q_k - Q^* = x_k + p_k = z_k + y_k + u_k + v_k.$$

Taking the infinity norm, using the triangle inequality, and then taking expectations gives

$$\mathbb{E}[\|Q_k - Q^*\|_\infty] \le \|z_k\|_\infty + \mathbb{E}[\|y_k\|_\infty] + \mathbb{E}[\|u_k\|_\infty] + \mathbb{E}[\|v_k\|_\infty].$$

Substituting Equations (80), (81), (83) and (84) into the preceding estimate yields

$$\begin{aligned}
\mathbb{E}[\|Q_k - Q^*\|_\infty] \le\ & K_{\beta_\varepsilon}\beta_\varepsilon^k\|Q_0 - Q^*\|_\infty \\
& + 2\alpha\gamma d_{\max}K_{\beta_\varepsilon}^2 k\beta_\varepsilon^{k-1}\|Q_0 - Q^*\|_\infty \\
& + 128|\mathcal{S}|^2|\mathcal{A}|^2\sqrt{W_{\max}}\left(t_{\mathrm{mix}}\sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)}\right) \\
& + \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)} 128|\mathcal{S}|^2|\mathcal{A}|^2\sqrt{W_{\max}}\left(t_{\mathrm{mix}}\sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)}\right).
\end{aligned}$$

The last two lines have the same Markovian-noise factor, so collecting them gives

$$\begin{aligned}
\mathbb{E}[\|Q_k - Q^*\|_\infty] \le\ & \left(1 + \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)}\right) 128|\mathcal{S}|^2|\mathcal{A}|^2\sqrt{W_{\max}} \\
& \times \left(t_{\mathrm{mix}}\sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)}\right) \\
& + K_{\beta_\varepsilon}\beta_\varepsilon^k\|Q_0 - Q^*\|_\infty \\
& + 2\alpha\gamma d_{\max}K_{\beta_\varepsilon}^2 k\beta_\varepsilon^{k-1}\|Q_0 - Q^*\|_\infty.
\end{aligned}$$

This is exactly Equation (10), with the convention $k\beta_\varepsilon^{k-1} = 0$ at $k = 0$. □

The weighted-norm certificate therefore gives the following Markovian finite-time bound.

Corollary 3. *Let Assumptions 1 and 2 hold, with d equal to the stationary distribution of the behavior-induced chain, and suppose that (β, q, v) satisfies Equation (8) with $\beta < 1$. Then, for every $k \ge 0$,*

$$\begin{aligned}
\mathbb{E}[\|Q_k - Q^*\|_\infty] \le\ & \left(1 + \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)}\right) 128|\mathcal{S}|^2|\mathcal{A}|^2\sqrt{W_{\max}} \\
& \times \left(t_{\mathrm{mix}}\sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)}\right) \\
& + \kappa_q\beta^k\|Q_0 - Q^*\|_\infty \\
& + 2\alpha\gamma d_{\max}\kappa_q^2 k\beta^{k-1}\|Q_0 - Q^*\|_\infty,
\end{aligned}$$

where $k\beta^{k-1} = 0$ when $k = 0$. In particular, if the infimum in Theorem 4 is attained, the same bound holds with $\beta = \rho(\mathcal{M}_\alpha)$.

Proof. The reference, residual, and Markovian-noise vectors and active modes satisfy

$$x_k, z_k, y_k, w_k, p_k, u_k, v_k, Q_k - Q^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, \qquad M_{\bar{\mu}}, M_{\mu_k} \in \mathbb{R}^{(|\mathcal{S}||\mathcal{A}|)\times(|\mathcal{S}||\mathcal{A}|)}.$$

Fix an arbitrary stochastic policy $\bar{\mu}$ and introduce the Markovian reference system

$$x_{k+1} = M_{\bar{\mu}}x_k + \alpha w_k, \qquad x_0 = Q_0 - Q^*.$$

Write $x_k = z_k + y_k$, where

$$z_{k+1} = M_{\bar{\mu}} z_k, \qquad z_0 = Q_0 - Q^*,$$
$$y_{k+1} = M_{\bar{\mu}} y_k + \alpha w_k, \quad y_0 = 0.$$

Because $M_{\bar{\mu}} \in \mathrm{co}(\mathcal{M}_\alpha)$, Equation (18) gives

$$\|z_k\|_\infty \le \kappa_q \beta^k \|Q_0 - Q^*\|_\infty. \tag{85}$$

The Poisson-equation analysis of the Markovian reference noise does not use the product-growth constant. Hence Lemma 32 continues to give

$$\mathbb{E}[\|y_k\|_\infty] \le 128|\mathcal{S}|^2|\mathcal{A}|^2\sqrt{W_{\max}} \times \left(t_{\mathrm{mix}}\sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)}\right). \tag{86}$$

Set $p_k := Q_k - Q^* - x_k$. Subtracting the reference system from the exact Markovian direct SLS in Equation (64) gives

$$p_{k+1} = M_{\mu_k} p_k + (M_{\mu_k} - M_{\bar{\mu}}) x_k, \qquad p_0 = 0.$$

The same row-wise comparison as in Equation (82) gives

$$\|M_{\mu_k} - M_{\bar{\mu}}\|_\infty \le 2\alpha\gamma d_{\max}. \tag{87}$$

Split $p_k = u_k + v_k$, where

$$u_{k+1} = M_{\mu_k} u_k + (M_{\mu_k} - M_{\bar{\mu}}) z_k, \quad u_0 = 0,$$
$$v_{k+1} = M_{\mu_k} v_k + (M_{\mu_k} - M_{\bar{\mu}}) y_k, \quad v_0 = 0.$$

For the deterministic residual, unrolling the recursion and using Equations (18), (85) and (87) yields

$$\begin{aligned}\|u_k\|_\infty &\le \sum_{i=0}^{k-1} \kappa_q \beta^{k-1-i} \cdot 2\alpha\gamma d_{\max} \cdot \kappa_q \beta^i \|Q_0 - Q^*\|_\infty \\ &= 2\alpha\gamma d_{\max} \kappa_q^2 k \beta^{k-1} \|Q_0 - Q^*\|_\infty.\end{aligned} \tag{88}$$

For the noise-driven residual, use the common row contraction and Equations (86) and (87). Then

$$\begin{aligned}\mathbb{E}[\|v_k\|_\infty] &\le \sum_{i=0}^{k-1} \rho_{\mathrm{row}}^{k-1-i} \cdot 2\alpha\gamma d_{\max} \cdot \mathbb{E}[\|y_i\|_\infty] \\ &\le 2\alpha\gamma d_{\max} 128|\mathcal{S}|^2|\mathcal{A}|^2\sqrt{W_{\max}} \\ &\quad \times \left(t_{\mathrm{mix}}\sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)}\right) \sum_{i=0}^{k-1} \rho_{\mathrm{row}}^{k-1-i} \\ &\le \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)} 128|\mathcal{S}|^2|\mathcal{A}|^2\sqrt{W_{\max}} \\ &\quad \times \left(t_{\mathrm{mix}}\sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)}\right),\end{aligned} \tag{89}$$

where the last inequality uses $1 - \rho_{\mathrm{row}} = \alpha d_{\min}(1-\gamma)$.

Finally,

$$Q_k - Q^* = z_k + y_k + u_k + v_k.$$

Combining the triangle inequality with Equations (85), (86), (88) and (89) gives

$$\begin{aligned}\mathbb{E}[\|Q_k - Q^*\|_\infty] \le &\left(1 + \frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)}\right) 128|\mathcal{S}|^2|\mathcal{A}|^2\sqrt{W_{\max}} \\ &\times \left(t_{\mathrm{mix}}\sqrt{\frac{\alpha}{d_{\min}(1-\gamma)}} + \frac{\alpha t_{\mathrm{mix}}}{d_{\min}(1-\gamma)}\right) \\ &+ \kappa_q \beta^k \|Q_0 - Q^*\|_\infty \\ &+ 2\alpha\gamma d_{\max} \kappa_q^2 k \beta^{k-1} \|Q_0 - Q^*\|_\infty,\end{aligned}$$

which is the claimed estimate. If the infimum in Theorem 4 is attained, choose a feasible pair (q, v) at $\beta = \rho(\mathcal{M}_\alpha)$; the same derivation then proves the final statement. □