BEYOND THE BELLMAN FIXED POINT: GEOMETRY AND FAST POLICY IDENTIFICATION IN VALUE ITERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Q-value iteration (Q-VI) is usually analyzed through the γ-contraction of the Bellman operator. This argument proves convergence to the optimal Q-function, Q^*, but it gives only a coarse account of when the induced greedy policy becomes optimal. We study discounted Q-VI as a switching system and focus on the practically optimal solution set (POSS), the set of Q-functions whose tie-broken greedy policies are optimal. The main result shows that Q-VI reaches the optimal action class in finite time by entering an invariant tube around $\mathcal{X}_1 = Q^* + \mathrm{span}(\mathbf{1})$, which is contained in the POSS, where $\mathbf{1}$ is the all-ones vector. For every $\varepsilon > 0$ such that $\bar{\rho} + \varepsilon \in (0, 1)$, the distance to $\mathcal{X}_1$ satisfies an exponential bound with rate $(\bar{\rho} + \varepsilon)^k$, where $\bar{\rho}$ is the joint spectral radius (JSR) of the projected switching family restricted to directions transverse to $\mathcal{X}_1$. When $\bar{\rho} < \gamma$, this transverse convergence is faster than the classical contraction rate. The analysis separates fast policy identification from the subsequent convergence to Q^*, which may be governed by the dynamics along the all-ones eigenvector direction. We also give spectral and graph-theoretic conditions under which the strict inequality $\bar{\rho} < \gamma$ holds or fails.

1 INTRODUCTION

Dynamic programming Bellman (1966); Bertsekas & Tsitsiklis (1996); Bertsekas (2012) is a standard method for solving Markov decision processes (MDPs) Puterman (1994). Among its basic algorithms, Q-value iteration (Q-VI) is notable for its simple update rule and its classical contraction guarantee Bertsekas & Tsitsiklis (1996); Bertsekas (2012). In a discounted problem with discount factor $\gamma \in (0, 1)$, the Bellman optimality operator is a γ-contraction in the infinity norm. Hence the Q-VI iterates converge exponentially to the optimal Q-function Q^* at rate γ, a result that underlies much of dynamic programming and reinforcement learning Sutton & Barto (1998).

The geometric mechanism considered in this paper is shown schematically in Figure 1. The usual contraction proof controls the distance to the fixed point Q^*, but it does not resolve the geometry of the trajectory in directions that do not affect action preferences. This distinction matters when the objective is to determine when the greedy policy has already become optimal, even though the Q-function itself has not yet converged. To make this question precise, we define the practically optimal solution set (POSS), denoted by $\mathcal{X}^*$, as the set of Q-functions whose tie-broken greedy policies are optimal.

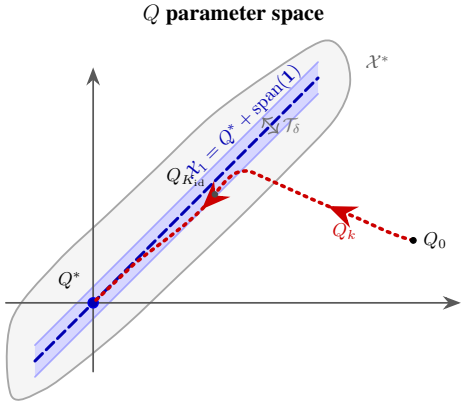


Figure 1: Geometric mechanism for finite-time POSS identification.

The starting point of the paper is a switching-system representation of Q-VI. The Bellman update can be written as an affine switching system Liberzon (2003); Lin & Antsaklis (2009); Hu et al. (2017), with the active mode selected by the tie-broken greedy pol-

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.

icy of the current iterate. This representation connects Q-VI with tools from switching-system theory Liberzon (2003); Lin & Antsaklis (2009); Hu et al. (2017) and Lyapunov analysis Khalil (2002). It also singles out the affine space

$$\mathcal{X}_1 := Q^* + \mathrm{span}(\mathbf{1}).$$

A shift of Q^* in the all-ones direction preserves every statewise action ordering, so $\mathcal{X}_1 \subset \mathcal{X}^*$. We prove that a sufficiently small tube around $\mathcal{X}_1$ is not only contained in $\mathcal{X}^*$, but also invariant under the Bellman operator. Once the iterates enter this tube, the tie-broken greedy policy is optimal, and the entrance time is finite.

The rate of this entrance is governed by the dynamics transverse to $\mathcal{X}_1$. The classical contraction argument treats the all-ones direction and the transverse directions together, giving the single rate γ. By deriving an exact switching linear system representation for the Q-VI error and projecting out the all-ones direction, we obtain a sharper bound: the distance from Q_k to $\mathcal{X}_1$ decays exponentially at any rate larger than the joint spectral radius (JSR) Tsitsiklis & Blondel (1997); Rota & Strang (1960); Blondel & Nesterov (2005) of the restricted projected switching system dynamics. If this restricted JSR is smaller than γ, then Q-VI approaches $\mathcal{X}_1$, and therefore identifies the POSS, faster than the standard fixed-point contraction rate would suggest. The resulting trajectory has two stages: rapid transverse approach and policy identification, followed by convergence to Q^*, possibly dominated by the remaining all-ones component.

We also study when the restricted JSR is strictly below the classical rate $\gamma \in (0, 1)$. Appendix B gives conditions for this strict inequality by describing when different state–action trajectories acquire overlapping future behavior after a uniformly bounded number of policy steps.

2 Related work

Switching system methods have recently been used to analyze reinforcement learning Sutton & Barto (1998) and dynamic programming Bellman (1966); Bertsekas & Tsitsiklis (1996); Bertsekas (2012). The framework of Lee & He (2020) gives a unified switching system view of Q-learning algorithms, and Lee et al. (2023) develops a discrete-time formulation with convergence guarantees. Subsequent work established sharper final iteration bounds Lee (2024) and finite-time results under Markovian observations and diminishing step-sizes Lim & Lee (2024).

For Q-VI, several papers have examined geometric or Lyapunov-based structure. A switching system interpretation and orthant-wise geometric behavior are studied in Lee (2023). Convex-analytic and Lyapunov methods for value-based reinforcement learning appear in Guo & Hu (2022), while Iervolino et al. (2023) formulates value iteration as an affine switching system. Switching-based policy-selection mechanisms are considered in Tipaldi et al. (2025). Goyal and Grand-Clément Goyal & Grand-Clément (2023) develop a first-order approach to accelerated value iteration and analyze value-iteration-type recursions through compositions of affine maps, or equivalently through a linear time-varying/switching dynamical-system viewpoint. These works are close in spirit to the present paper, as they use dynamical-systems tools to study value-based dynamic programming and reinforcement learning algorithms.

Geometric approaches outside switching systems provide a complementary perspective. The value-function polytope framework of Dadashi et al. (2019) characterizes geometric and topological properties of value functions in finite MDPs and uses that structure to interpret reinforcement learning algorithms. Related polyhedral ideas are used in Wu & De Loera (2022), which develops geometric policy iteration through hyperplane arrangements and boundary structures of MDPs. Those works focus mainly on value-function geometry, policy improvement, and polyhedral structure. Our focus is instead on the Q-VI trajectory relative to $\mathcal{X}_1 = Q^* + \mathrm{span}(\mathbf{1})$ and the POSS $\mathcal{X}^*$.

Finite-time policy identification is also related to turnpike results for discounted finite MDPs. Feinberg and He Feinberg & He (2025) study turnpike functions and turnpike integers, showing that, after a finite number of value-iteration steps, the generated decision rules are deterministic optimal policies for the infinite-horizon problem. This is close in spirit to finite identification, but their analysis is formulated through turnpike functions for VI, whereas our results study Q-VI trajectory geometry, invariant tubes around $\mathcal{X}_1$, and projected switching dynamics.

Other analyses of Q-VI rely on probabilistic or nonsmooth viewpoints. The works Mustafin et al. (2024; 2025a;b) use tools such as absolute probability sequences to refine the convergence analysis of classical MDP solution methods. Semismooth Newton-type interpretations of dynamic programming are studied in Gargiani et al. (2022). These nonsmooth-equation perspectives are relevant because Bellman optimality operators contain maximization and policy selection. Here we use the same nonsmooth structure to obtain switching dynamics and finite-time identification of the optimal action class.

The technical tools are also connected to stability theory for switching systems and to JSR analysis. The JSR was introduced in Rota & Strang (1960), and Lyapunov methods for bounding or approximating it have been extensively developed; see, for example, Tsitsiklis & Blondel (1997). Our Lyapunov construction follows this general line but is applied after projecting away the all-ones direction. The projection is also related to seminorm-based analyses in reinforcement learning and dynamic programming. In particular, the non-asymptotic seminorm Lyapunov framework of Chen et al. (2025) studies deterministic and stochastic iterative algorithms through seminorm contractions.

Most existing results emphasize asymptotic convergence, global contraction, or the geometry of value functions and policies. The present paper instead analyzes the geometry of the Q-VI trajectory itself, with an emphasis on finite-time policy identification, invariant tubes, and projected switching system dynamics.

3 Preliminaries

3.1 Notations

We use the following notation. The symbols $\mathbb{R}$, $\mathbb{R}^n$, and $\mathbb{R}^{n\times m}$ denote the set of real numbers, the n-dimensional Euclidean space, and the set of $n \times m$ real matrices, respectively. For a matrix A, $A^\top$ denotes its transpose. The relations $A \succ 0$, $A \prec 0$, $A \succeq 0$, and $A \preceq 0$ mean that A is symmetric positive definite, negative definite, positive semidefinite, and negative semidefinite, respectively. The identity matrix with appropriate dimensions is denoted by I. For a finite set $\mathcal{S}$, its cardinality is denoted by $|\mathcal{S}|$. The Kronecker product of A and B is denoted by $A \otimes B$. The vector with all entries equal to one is denoted by $\mathbf{1}$, and $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the maximum and minimum eigenvalues, respectively.

For a convex set $\mathcal{Y} \subset \mathbb{R}^n$ and a vector $x \in \mathbb{R}^n$, $\mathrm{dist}_2(x, \mathcal{Y})$ and $\mathrm{dist}_\infty(x, \mathcal{Y})$ denote the Euclidean distance and the infinity-norm distance from x to $\mathcal{Y}$, respectively:

$$\mathrm{dist}_2(x, \mathcal{Y}) := \inf_{y\in\mathcal{Y}} \|x - y\|_2, \qquad \mathrm{dist}_\infty(x, \mathcal{Y}) := \inf_{y\in\mathcal{Y}} \|x - y\|_\infty.$$

We write $\Delta_{|\mathcal{A}|}$ for the probability simplex over a finite action set $\mathcal{A}$:

$$\Delta_{|\mathcal{A}|} := \left\{ p \in \mathbb{R}^{|\mathcal{A}|} : \; p_i \geq 0, \; \sum_{i=1}^{|\mathcal{A}|} p_i = 1 \right\}.$$

Throughout the paper, $\mathrm{Arg\,max}$ denotes the set-valued maximizer, while $\arg\max$ denotes a fixed tie-broken single-valued maximizer.

3.2 Markov decision process

We consider an infinite-horizon discounted Markov decision process (MDP) Puterman (1994), in which an agent sequentially chooses actions to maximize cumulative discounted rewards. The state and action spaces are finite and are denoted by $\mathcal{S} := \{1, 2, \ldots, |\mathcal{S}|\}$ and $\mathcal{A} := \{1, 2, \ldots, |\mathcal{A}|\}$, respectively. At state $s \in \mathcal{S}$, the decision maker selects an action $a \in \mathcal{A}$. The next state s' is drawn according to $P(s'|s, a)$, and the transition incurs reward $r(s, a, s')$, where $r : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$. We write $r(s_k, a_k, s_{k+1}) =: r_{k+1}$ for $k \geq 0$. A deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$ maps each state s to an action $\pi(s)$. Throughout the paper, the discount factor satisfies $\gamma \in (0, 1)$.

For a policy π, the Q-function under π is defined as

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{k+1} \,\middle|\, s_0 = s, a_0 = a, \pi \right],$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. The optimal Q-function is

$$Q^*(s,a) := \sup_{\pi \in \Theta} Q^\pi(s,a), \qquad s \in \mathcal{S},\ a \in \mathcal{A},$$

where Θ denotes the set of all admissible deterministic policies. Define $V^\pi(s) := Q^\pi(s, \pi(s))$ and $V^*(s) := \max_{a \in \mathcal{A}} Q^*(s,a)$. A deterministic policy π^* is optimal if $V^{\pi^*}(s) = V^*(s)$ for all $s \in \mathcal{S}$. Once Q^* is known, an optimal tie-broken greedy policy can be recovered as $\pi^*(s) = \arg\max_{a \in \mathcal{A}} Q^*(s,a)$.

For each state $s \in \mathcal{S}$, define the set of optimal greedy actions by

$$\Phi^*(s) := \operatorname{Arg\,max}_{a \in \mathcal{A}} Q^*(s,a).$$

The set of all optimal deterministic policies is then

$$\Theta^* := \{\pi \in \Theta :\ \pi(s) \in \Phi^*(s),\ \forall s \in \mathcal{S}\}.$$

For completeness, we recall why this set coincides with the set of optimal deterministic policies in a finite discounted MDP. If $\pi \in \Theta^*$, then

$$V^*(s) = Q^*(s,\pi(s)) = R(s,\pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,\pi(s)) V^*(s'), \qquad \forall s \in \mathcal{S}.$$

Therefore, V^* is a fixed point of the discounted Bellman evaluation operator for π. By uniqueness of this fixed point, $V^\pi = V^*$, so π is optimal. Consequently,

$$Q^\pi(s,a) = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^\pi(s') = R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) V^*(s') = Q^*(s,a).$$

Conversely, if a deterministic policy π is optimal, then $V^\pi = V^*$ by definition. Hence

$$\begin{aligned} V^*(s) =& V^\pi(s) = R(s,\pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,\pi(s)) V^\pi(s') \\ =& R(s,\pi(s)) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,\pi(s)) V^*(s') = Q^*(s,\pi(s)), \end{aligned}$$

and hence, $\pi(s) \in \Phi^*(s)$ for all states. Therefore, the preceding characterization describes all optimal deterministic policies.

3.3 Q-value iteration

We study Q-value iteration (Q-VI) Bertsekas & Tsitsiklis (1996). Let F denote the Bellman operator

$$(FQ)(s,a) := R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{a' \in \mathcal{A}} Q(s',a'), \quad (s,a) \in \mathcal{S} \times \mathcal{A}.$$

The Q-VI recursion is

$$Q_{k+1} = FQ_k := F(Q_k), \qquad k = 0, 1, \ldots. \tag{1}$$

The corresponding pseudocode is given in Appendix A.1 as Algorithm 1. It is well known that Q-VI converges exponentially to Q^* in the infinity norm $\|\cdot\|_\infty$ (Bertsekas & Tsitsiklis, 1996, Lemma 2.5).

Lemma 1 (Bertsekas & Tsitsiklis (1996))**.** *The Q-VI iterates satisfy* $\|Q_{k+1} - Q^*\|_\infty \le \gamma \|Q_k - Q^*\|_\infty$.

The proof in (Bertsekas & Tsitsiklis, 1996, Lemma 2.5) is based on the contraction property of the Bellman operator. A direct consequence of Lemma 1 is

$$\|Q_k - Q^*\|_\infty \le \gamma^k \|Q_0 - Q^*\|_\infty . \tag{2}$$

The standard estimate in Equation (2) controls convergence to the fixed point Q^*. The rest of the paper studies the geometry of the same recursion relative to the affine set $\mathcal{X}_1$ and the POSS $\mathcal{X}^*$.

4 Switching system model of Q-VI

This section introduces the switching system terminology and the compact notation used later. The full affine switching derivation is given in Appendix A.

4.1 SWITCHING SYSTEM

Let us consider the switching linear system Liberzon (2003); Lin & Antsaklis (2009)

$$x_{k+1} = A_{\sigma_k} x_k, \quad x_0 = z \in \mathbb{R}^n, \quad k \in \{0, 1, \ldots\},$$

where $x_k \in \mathbb{R}^n$ is the state, $\sigma_k \in \mathcal{M} := \{1, 2, \ldots, M\}$ is the switching signal, and $\mathcal{H} := \{A_1, A_2, \ldots, A_M\}$ is the switching family of subsystem matrices. Linear switching systems and their control theory are well studied Liberzon (2003); Lin & Antsaklis (2009). The joint spectral radius (JSR) Tsitsiklis & Blondel (1997); Rota & Strang (1960); Blondel & Nesterov (2005) measures the largest exponential growth rate attainable under arbitrary switching among the subsystem matrices.

Definition 1 (Joint spectral radius Rota & Strang (1960))**.** *Given a set of matrices $\{A_i \in \mathbb{R}^{n\times n}\}_{i=1}^M$, the JSR is defined as*

$$\rho(A_1, \cdots, A_M) = \lim_{k\to\infty} \max_{\bar{\sigma}_k \in \mathcal{M}^k} \|A_{\sigma_k} \cdots A_{\sigma_2} A_{\sigma_1}\|^{1/k},$$

where $\|\cdot\|$ is any submultiplicative matrix norm, and $\bar{\sigma}_k := (\sigma_1, \sigma_2, \ldots, \sigma_k) \in \mathcal{M}^k$.

The JSR is independent of the chosen submultiplicative norm. A related class of systems is the affine switching system

$$x_{k+1} = A_{\sigma_k} x_k + b_{\sigma_k}, \quad x_0 = z \in \mathbb{R}^n, \quad k \in \{0, 1, \ldots\},$$

where the input vector $b_{\sigma_k} \in \mathbb{R}^n$ also switches according to σ_k. The affine term introduces a dependence on the active mode that is absent from purely linear switching systems.

4.2 DEFINITIONS

Throughout the paper, we use the following compact notation:

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

where $P_a = P(\cdot|\cdot, a) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$, $Q(\cdot, a) \in \mathbb{R}^{|\mathcal{S}|}$ for $a \in \mathcal{A}$, and $R \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ enumerates $R(s, a) := \mathbb{E}[r_{k+1}|s_k = s, a_k = a]$ in an order compatible with the vectorization of Q. Thus, a Q-function is encoded as a single vector $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ whose entries list $Q(s, a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. In particular, $Q(s, a) = (e_a \otimes e_s)^\top Q$, where $e_s \in \mathbb{R}^{|\mathcal{S}|}$ and $e_a \in \mathbb{R}^{|\mathcal{A}|}$ are the s-th and a-th basis vectors, respectively. For any stochastic policy $\pi : \mathcal{S} \to \Delta_{|\mathcal{A}|}$, define the corresponding action transition matrix by

$$\Pi^\pi := \begin{bmatrix} \pi(1)^\top \otimes e_1^\top \\ \pi(2)^\top \otimes e_2^\top \\ \vdots \\ \pi(|\mathcal{S}|)^\top \otimes e_{|\mathcal{S}|}^\top \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}, \tag{3}$$

where $e_s \in \mathbb{R}^{|\mathcal{S}|}$. Then $P\Pi^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ is the transition probability matrix of the state-action pair under policy π. For a deterministic policy $\pi : \mathcal{S} \to \mathcal{A}$, we use the corresponding one-hot vector $\vec{\pi}(s) := e_{\pi(s)} \in \Delta_{|\mathcal{A}|}$, and the action transition matrix is given by Equation (3) with π replaced by $\vec{\pi}$. For any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, let $\pi_Q(s) := \arg\max_{a\in\mathcal{A}} Q(s, a)$ denote the tie-broken greedy policy with respect to Q. We also use the shorthand

$$\Pi_Q := \Pi^{\pi_Q}.$$

4.3 SWITCHING-SYSTEM REPRESENTATION OF Q-VI

With this vector notation, Q-VI can be naturally written as an affine switching system without altering the recursion. In compact form, the Bellman update is

$$Q_{k+1} = R + \gamma P \Pi_{Q_k} Q_k = F Q_k,$$

where the active matrix is determined by the tie-broken greedy policy induced by the current iterate. For each deterministic policy $\pi_i \in \Theta$, write

$$A_i := \gamma P \Pi^{\pi_i}, \qquad i \in \{1, \dots, |\Theta|\}.$$

The detailed switching affine system representation of the error, together with the Q-VI algorithm and the basic norm and JSR facts for the full switching family, is given in Appendix A. This formulation supplies the dynamical-systems notation used below. The geometric analysis begins with the invariant tube around $\mathcal{X}_1$.

5 Invariant tube and practically optimal solution set

The switching system viewpoint summarized above, and detailed in Appendix A, gives a dynamical framework for Q-VI trajectories. The analysis rests on two geometric objects: the practically optimal solution set (POSS) and an invariant tube around a distinguished affine subspace. Together they formalize the fact that Q-VI may identify the optimal action class before it reaches Q^*. The POSS is defined by

$$\mathcal{X}^* := \left\{ Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} : \ \pi_Q(s) \in \Phi^*(s), \ \forall s \in \mathcal{S} \right\},$$

where π_Q is the tie-broken greedy policy induced by Q:

$$\pi_Q(s) := \arg\max_{a \in \mathcal{A}} Q(s, a).$$

Thus, $Q \in \mathcal{X}^*$ means that the induced tie-broken greedy policy π_Q is optimal, even if $Q \neq Q^*$. We also define the affine space

$$\mathcal{X}_1 := \left\{ Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} : Q = Q^* + \alpha \mathbf{1}, \ \alpha \in \mathbb{R} \right\} = \operatorname{span}(\mathbf{1}) + Q^*.$$

Adding a constant $\alpha \mathbf{1}$ to Q^* does not change action orderings. Therefore, the resulting tie-broken greedy policy remains optimal. Hence $\mathcal{X}_1 \subset \mathcal{X}^*$; see Lemma 4 in Appendix D. The affine set $\mathcal{X}_1$ plays a central role because it is invariant under the Bellman operator F. This invariance property is stated and proved in Proposition 1 in Appendix D.

It remains to ask whether $\mathcal{X}^*$ itself is invariant under F. In general it is not. However, a sufficiently small tube around $\mathcal{X}_1 = Q^* + \operatorname{span}(\mathbf{1})$ is both invariant and contained in $\mathcal{X}^*$. This link between set invariance and policy identification is the key point of the section. We first exclude the degenerate case in which every action is optimal at every state.

Assumption 1 (Optimal-class separation)**.** *Let*

$$\mathcal{S}_{\text{sep}} := \{ s \in \mathcal{S} : \ \Phi^*(s) \neq \mathcal{A} \}.$$

We assume throughout that $\mathcal{S}_{\text{sep}} \neq \varnothing$.

Assumption 1 guarantees that at least one state has a strict separation between optimal and non-optimal actions. For each $s \in \mathcal{S}_{\text{sep}}$, define

$$\bar{\Delta}_s := V^*(s) - \max_{a \notin \Phi^*(s)} Q^*(s, a),$$

and let

$$\bar{\Delta} := \min_{s \in \mathcal{S}_{\text{sep}}} \bar{\Delta}_s.$$

Since the action space is finite, $\bar{\Delta} > 0$. If $\mathcal{S}_{\text{sep}} = \varnothing$, then every action is optimal at every state, and the identification problem is trivial. The assumption is imposed only to exclude this case. Next, we define the invariant tube as follows.

Proposition 2 (Invariant tube inside $\mathcal{X}^*$)**.** *Let Assumption 1 hold, and fix any* $\delta \in (0, \bar{\Delta}/2)$. *Define the tube around* $\mathcal{X}_1$ *by*

$$\mathcal{T}_\delta := \left\{ Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} : \ \operatorname{dist}_\infty(Q, \mathcal{X}_1) \leq \delta \right\}.$$

Then $\mathcal{T}_\delta \subset \mathcal{X}^*$. *Moreover,*

$$F(\mathcal{T}_\delta) \subset \mathcal{T}_{\gamma\delta} \subset \mathcal{T}_\delta,$$

so $\mathcal{T}_\delta$ *is invariant under the Bellman operator* F.

Proof. See Appendix D.3.

Consequently, once Q_k enters $\mathcal{T}_\delta$, every later Q-VI iterate remains in the tube. Because $\mathcal{T}_\delta \subset \mathcal{X}^*$, tube entrance also means POSS entrance. A basic finite-time entrance bound based on the standard γ-contraction is given in Corollary 1 in Appendix D.

Thus Q-VI need not reach Q^* in finite time in order for its tie-broken greedy policy to become optimal. In this sense it has a finite policy-identification property analogous to policy iteration.

The remaining issue is quantitative. Q-VI converges to Q^* at the classical rate γ, but the approach to the invariant tube can be faster. The next section makes this precise using a restricted JSR.

The mechanism in Figure 1 underlies both the invariant tube argument and the faster identification result. The Q-VI trajectory first moves toward $\mathcal{X}_1 = Q^* + \mathrm{span}(\mathbf{1})$, enters an invariant tube $\mathcal{T}_\delta \subset \mathcal{X}^*$, and thereby identifies an optimal tie-broken greedy policy in finite time. After identification, the remaining evolution continues toward Q^*, possibly along a slower dynamics associated with the all-ones direction.

6 Faster identification of the POSS

We next estimate how quickly Q-VI approaches the tube $\mathcal{T}_\delta$. Since $\mathcal{X}_1$ is the center line of this tube, it suffices to analyze how quickly the iterates approach $\mathcal{X}_1$. Although the tube is defined using dist_∞, the Euclidean distance used below is sufficient for tube entrance because

$$\mathrm{dist}_\infty(Q, \mathcal{X}_1) \le \mathrm{dist}_2(Q, \mathcal{X}_1), \qquad \forall Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}.$$

Thus, a sufficiently small ℓ_2-distance to $\mathcal{X}_1$ implies membership in the corresponding ℓ_∞-tube. The first step is an exact switching system representation of the Q-VI error over stochastic policies. The main idea behind the next lemma is adapted from the affine-map and linear time-varying viewpoint of Goyal and Grand-Clément Goyal & Grand-Clément (2023): the Bellman maximization can be represented by selecting maps from a switching system family, and products of those maps can then be analyzed dynamically. Here, we use the same idea for Q-VI, but represent the Bellman max-difference exactly through a state-dependent stochastic policy and then project away the all-ones direction.

Lemma 3 (Exact stochastic-policy representation of the Q-VI error)**.** *Let*

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

Then, for each $k \ge 0$, there exists a stochastic policy $\mu_k : \mathcal{S} \to \Delta_{|\mathcal{A}|}$ such that

$$e_{k+1} = A_{\mu_k} e_k,$$

where $A_\mu := \gamma P \Pi^\mu \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$.

Proof. See Appendix D.5.

The stochastic policy μ_k in Lemma 3 is only an auxiliary representation of the Bellman max-difference as a convex combination of current errors. It is generally not the greedy policy π_{Q_k}, need not be unique, and is not implemented by the algorithm.

In conclusion, the Q-VI error admits an exact linear switching system representation; no approximation is involved. The maximization in the Bellman operator is absorbed into the state-dependent stochastic policy μ_k. This permits the use of spectral and Lyapunov tools for the linear maps A_μ when studying convergence toward $\mathcal{X}_1$.

Because our main object is the distance from Q_k to

$$\mathcal{X}_1 = Q^* + \mathrm{span}(\mathbf{1}),$$

we isolate the component of the error orthogonal to $\mathrm{span}(\mathbf{1})$. The component along $\mathbf{1}$ only shifts the Q-function uniformly and does not affect the tie-broken greedy policy. Let

$$\mathbf{\Pi}_\perp := I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \qquad n := |\mathcal{S}||\mathcal{A}|,$$

be the orthogonal projection onto $\mathrm{span}(\mathbf{1})^{\perp}$. Define the projected error

$$z_k := \mathbf{\Pi}_{\perp} e_k \in \mathbb{R}^n,$$

the restricted matrices

$$\bar{A}_i = \mathbf{\Pi}_{\perp} A_i \mathbf{\Pi}_{\perp}, \quad i \in \{1, 2, \ldots, M\},$$

and, for each stochastic policy μ,

$$\bar{A}_\mu := \mathbf{\Pi}_{\perp} A_\mu \mathbf{\Pi}_{\perp}.$$

We denote the restricted switching family associated with deterministic policies and its JSR by

$$\bar{\mathcal{H}} := \{\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_M\}$$

and

$$\bar{\rho} := \rho(\bar{A}_1, \bar{A}_2, \ldots, \bar{A}_M),$$

respectively. The projected error then evolves under the restricted switching family associated with stochastic policies, and its norm is exactly the distance to $\mathcal{X}_1$.

Lemma 4 (Projected error dynamics)**.** *The projected error z_k satisfies*

$$z_{k+1} = \bar{A}_{\mu_k} z_k \qquad \forall k \geq 0.$$

Moreover, $\mathrm{dist}_2(Q_k, \mathcal{X}_1) = \|z_k\|_2$.

Proof. See Appendix D.6.

Several auxiliary facts about the restricted matrices are needed to analyze the projected dynamics in Lemma 4. Specifically, Lemma 7 in Appendix D explains how the projection removes the all-ones direction, Lemma 8 in Appendix D gives the corresponding product identity, Lemma 9 in Appendix D characterizes the spectral radius of each restricted subsystem, and Lemma 10 in Appendix D gives a norm-based computable upper bound on the restricted JSR.

Lemma 5. *The JSR of the restricted switching system satisfies*

$$\rho(\bar{A}_1, \ldots, \bar{A}_M) \leq \gamma.$$

Proof. See Appendix D.11.

Hence, the restricted JSR associated with the restricted projected switching family never exceeds the classical rate γ. An explicit exponential estimate follows from a common Lyapunov norm for the restricted switching family, together with an extension of the Lyapunov inequality to the convex hull. The first ingredient is stated and proved in Lemma 12 in Appendix D, and the second is stated and proved in Lemma 13 in Appendix D. Together, they allow the Lyapunov inequality to be applied directly to the stochastic-policy projected dynamics.

This gives exponential convergence of the actual Q-VI iterates toward $\mathcal{X}_1$.

Theorem 1 (Global exponential convergence to $\mathcal{X}_1$ for the actual Q-VI)**.** *Fix any $\epsilon > 0$ such that*

$$\beta_\epsilon := \bar{\rho} + \epsilon \in (0, 1).$$

There exists a constant $c_\varepsilon > 0$, depending on $\varepsilon > 0$, such that

$$\mathrm{dist}_2(Q_k, \mathcal{X}_1) \leq c_\varepsilon \beta_\varepsilon^k \, \mathrm{dist}_2(Q_0, \mathcal{X}_1), \qquad \forall k \geq 0. \tag{4}$$

Proof. See Appendix D.14.

This theorem gives exponential convergence to $\mathcal{X}_1$. Since $\bar{\rho} \leq \gamma$, the case $\bar{\rho} < \gamma$ permits a choice of $\varepsilon > 0$ with $\beta_\varepsilon < \gamma$. The transverse approach to $\mathcal{X}_1$ is then strictly faster than the standard Q-VI rate for convergence to Q^*.

The same estimate gives an explicit finite-time bound for entrance into the POSS; see Corollary 2 in Appendix D. If $\bar{\rho} < \gamma$, then ε can be chosen so that $\beta_\varepsilon < \gamma$, and the asymptotic transverse rate is strictly faster than the classical γ-rate. The resulting finite-time bound may be smaller than the basic γ-based estimate, depending on the constant c_ε and the initial transverse distance $\mathrm{dist}_2(Q_0, \mathcal{X}_1)$.

Corollary 2 in Appendix D gives finite-time identification of the POSS through convergence toward $\mathcal{X}_1$. Once the iterate is close enough to $\mathcal{X}_1$, its tie-broken greedy policy selects only optimal actions, even though the full Q-function has not yet converged to Q^*. When $\beta_\varepsilon < \gamma$, this provides a potentially sharper policy-identification estimate than the classical contraction argument.

The geometry is therefore two-stage. First, Q_k approaches $\mathcal{X}_1$, enters $\mathcal{T}_\delta$, and hence enters $\mathcal{X}^*$. From then on, every tie-broken greedy policy π_{Q_k} is optimal, so the remaining iterations amount to evaluation over optimal policies. The final convergence to Q^* may still be limited by the standard γ-rate.

7 Two-stage convergence

The finite-time identification bound in Corollary 2 in Appendix D gives a two-stage description of the Q-VI trajectory. The full statement is given in Theorem 2 in Appendix D. Before identification, the transverse distance to $\mathcal{X}_1$ admits exponential upper bounds at any rate larger than the full restricted JSR $\bar{\rho}$. After identification, only optimal policies remain active, and the transverse component admits analogous bounds at any rate larger than the JSR $\bar{\rho}_*$ of the smaller optimal restricted family. If $\bar{\rho}_* < \gamma$, then the post-identification transverse decay is strictly faster than the standard γ-rate, although the full error may still be dominated by the residual all-ones component.

When the optimal policy is unique, the post-identification switching disappears. This sharper special case is stated in Corollary 3 in Appendix D; it reduces the second stage to a single linear system associated with π^*, whose transverse rate is $\rho(\bar{A}_{\pi^*}) = \gamma|\lambda_2(P\Pi^{\pi^*})|$.

The numerical examples below show the mechanism for deterministic Q-VI and for tabular Q-learning. The Q-VI experiment demonstrates finite-time entrance into the POSS through approach to $\mathcal{X}_1$, while the Q-learning experiment suggests that the same geometric pattern can persist under stochastic updates.

Remark 1 (When the transverse rate is strictly faster)**.** *Appendix B gives conditions ensuring $\bar{\rho} < \gamma$; see Proposition 3 in Appendix B. Intuitively, the strict inequality holds when every sufficiently long sequence of policy steps makes trajectories starting from different state–action pairs share some possible future behavior. This overlap makes differences that affect action preferences shrink faster than the classical γ-rate.*

8 Examples

This section gives a short summary of the numerical examples; the complete MDP specification, numerical values, figures, and discussion are in Appendix C. The appendix considers a small discounted MDP with $\mathcal{S} = \{1, 2, 3\}$, $\mathcal{A} = \{1, 2\}$, and $\gamma = 0.95$; for this MDP, the unique optimal deterministic policy is $\pi^* = \big(\pi^*(1), \pi^*(2), \pi^*(3)\big) = (1, 1, 1)$. It also includes a tabular Q-learning experiment on the same problem. See Appendix C, especially Figures 2 to 5, for the corresponding plots.

For Q-VI, the normalized decay plot in Figure 2 shows that $\mathrm{dist}_2(Q_k, \mathcal{X}_1)$ decreases faster than $\|Q_k - Q^*\|_\infty$ in the transient regime. The projected single-trajectory and multi-trajectory plots in Figures 3 and 4 show further that the iterates are rapidly drawn toward the affine set $\mathcal{X}_1$, enter the tube $\mathcal{T}_\delta$ in finite time, and then converge to Q^*. The single Q-VI trajectory shown is used to visualize tube entrance; its initial tie-broken greedy policy is already optimal, so it should not be interpreted as a first-time policy-identification example from a non-optimal greedy policy. For tabular Q-learning, Figure 5 shows the analogous stochastic behavior: despite sample-path fluctuations, trajectories from different initial points tend to approach the neighborhood of $\mathcal{X}_1$ first and only afterward refine their convergence toward Q^*.

9 Conclusion

We studied discounted Q-VI from a switching-system perspective and proved that Q-VI identifies the optimal action class in finite time, before asymptotic convergence to Q^* is necessarily complete. The mechanism is the approach to the affine set $\mathcal{X}_1 = Q^* + \mathrm{span}(\mathbf{1})$ and entrance into an invariant

tube contained in the POSS. This approach admits exponential upper bounds at any rate larger than the restricted JSR and can be faster than the classical γ-rate when the restricted JSR is strictly smaller than γ. We also gave structural conditions, based on spectral and graph-theoretic arguments, that explain when this strict improvement occurs and when Markov-chain obstructions prevent it. The resulting two-stage description separates fast policy identification from final convergence to the Bellman fixed point.

Reproducibility Statement

This paper primarily presents theoretical contributions. The main text states the problem setting and assumptions, and Appendix D provides detailed proofs and auxiliary results. The illustrative examples and numerical simulations are intentionally simple and can be readily reconstructed from the explicit problem specifications and parameter settings in Appendix C. The accompanying Python script and notebook in the supplementary materials provide implementations of these examples to facilitate reproduction of the numerical results.

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

Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

Dimitri P. Bertsekas. *Dynamic Programming and Optimal Control, Volume II*. Athena Scientific, Belmont, MA, 4 edition, 2012.

Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA, 1996.

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.

Zaiwei Chen, Sheng Zhang, Zhe Zhang, Shaan Ul Haque, and Siva Theja Maguluri. A non-asymptotic theory of seminorm lyapunov stability: From deterministic to stochastic iterative algorithms. *arXiv preprint arXiv:2502.14208*, 2025.

Robert Dadashi, Adrien Ali Taiga, Nicolas Le Roux, Dale Schuurmans, and Marc G. Bellemare. The value function polytope in reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1486–1495. PMLR, 2019.

R. L. Dobrushin. Central limit theorem for nonstationary Markov chains. I. *Theory of Probability and Its Applications*, 1(1):65–80, 1956.

Eugene A. Feinberg and Gaojin He. Properties of turnpike functions for discounted Markov decision processes. In *2025 Proceedings of the Conference on Control and Its Applications*, pp. 23–30. Society for Industrial and Applied Mathematics, 2025. doi:10.1137/1.9781611978742.4.

Matilde Gargiani, Andrea Zanelli, Dominic Liao-McPherson, Tyler H Summers, and John Lygeros. Dynamic programming through the lens of semismooth Newton-type methods. *IEEE Control Systems Letters*, 6:2996–3001, 2022.

Stéphane Gaubert and Zheng Qu. Dobrushin's ergodicity coefficient for Markov operators on cones. *Integral Equations and Operator Theory*, 81(1):127–150, 2015.

Vineet Goyal and Julien Grand-Clément. A first-order approach to accelerated value iteration. *Operations Research*, 71(2):517–535, 2023.

Xingang Guo and Bin Hu. Convex programs and Lyapunov functions for reinforcement learning: A unified perspective on the analysis of value-based methods. In *2022 American Control Conference (ACC)*, pp. 3317–3322, 2022.

John Hajnal and Maurice S. Bartlett. Weak ergodicity in non-homogeneous Markov chains. *Mathematical Proceedings of the Cambridge Philosophical Society*, 54(2):233–246, 1958.

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.

Raffaele Iervolino, Massimo Tipaldi, and Ali Forootani. A Lyapunov-based version of the value iteration algorithm formulated as a discrete-time switched affine system. *International Journal of Control*, 96(3):577–592, 2023.

Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, Upper Saddle River, NJ, 3 edition, 2002.

Donghwan Lee. On some geometric behavior of value iteration on the orthant: Switching system perspective. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 4911–4916, 2023.

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

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

David A. Levin and Yuval Peres. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, RI, 2 edition, 2017.

Daniel Liberzon. *Switching in Systems and Control*. Birkhäuser, Boston, MA, 2003.

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.

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.

Arsenii Mustafin, Alex Olshevsky, and Ioannis Ch Paschalidis. On value iteration convergence in connected MDPs. *arXiv preprint arXiv:2406.09592*, 2024.

Arsenii Mustafin, Sebastien Colla, Alex Olshevsky, and Ioannis Ch Paschalidis. Analysis of value iteration through absolute probability sequences. *arXiv preprint arXiv:2502.03244*, 2025a.

Arsenii Mustafin, Aleksei Pakharev, Alex Olshevsky, and Ioannis Ch Paschalidis. Geometric reanalysis of classical MDP solving algorithms. *arXiv preprint arXiv:2503.04203*, 2025b.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, New York, NY, 1994. doi:10.1002/9780470316887.

Gian-Carlo Rota and Gilbert Strang. A note on the joint spectral radius. *Indagationes Mathematicae*, 22:379–381, 1960.

Eugene Seneta. *Non-Negative Matrices and Markov Chains*. Springer, New York, 2006.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

Massimo Tipaldi, Raffaele Iervolino, Paolo Roberto Massenio, and David Naso. A switching control strategy for policy selection in stochastic dynamic programming problems. *Automatica*, 171: 111884, 2025.

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.

John N. Tsitsiklis, Dimitri P. Bertsekas, and Michael Athans. Distributed asynchronous deterministic and stochastic gradient optimization algorithms. *IEEE Transactions on Automatic Control*, 31(9): 803–812, 1986.

Jacob Wolfowitz. Products of indecomposable, aperiodic, stochastic matrices. *Proceedings of the American Mathematical Society*, 14(5):733–737, 1963.

Yue Wu and Jesús A. De Loera. Geometric policy iteration for Markov decision processes. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2070–2078, 2022. doi:10.1145/3534678.3539478.

A Q-VI ALGORITHM AND SWITCHING-SYSTEM REPRESENTATION

This appendix gives the algorithmic and switching-system details used in the main text. The subsequent derivation shows how the same update can be written as an affine switching system whose active mode is determined by the tie-broken greedy policy.

A.1 Q-VI ALGORITHM

Algorithm 1 Q-VI

1: Initialize $Q_0 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ arbitrarily.
2: **for** iteration $k = 0, 1, \ldots$ **do**
3: Update

$$Q_{k+1}(s,a) = \underbrace{R(s,a) + \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \max_{a' \in \mathcal{A}} Q_k(s',a')}_{=:FQ_k}$$

4: **end for**

Algorithm 1 is exactly the componentwise form of $Q_{k+1} = FQ_k$ in Equation (1). The componentwise form above makes explicit how the maximization over next-state actions enters the Bellman update.

A.2 SWITCHING-SYSTEM MODEL OF Q-VI

With the notation and the Bellman update in place, we reformulate the iteration as a switching system. The reformulation makes explicit that the iterate is governed by affine subsystems whose active mode is determined by the tie-broken greedy policy of the current Q-function.

Using the notation introduced in the main text, the update in Algorithm 1 can be rewritten as

$$Q_{k+1} = R + \gamma P \Pi_{Q_k} Q_k := F(Q_k). \tag{5}$$

Recall the definitions of $\pi_Q(s)$ and Π_Q. Invoking the optimal Bellman equation $(\gamma P \Pi_{Q^*} - I)Q^* + R = 0$, Equation (5) can be further rewritten as

$$(Q_{k+1} - Q^*) = \gamma P \Pi_{Q_k}(Q_k - Q^*) + \gamma P(\Pi_{Q_k} - \Pi_{Q^*})Q^*.$$

Thus, in error coordinates, Q-VI is an affine switching system. In particular, for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, define

$$A_Q := \gamma P \Pi_Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}, \quad b_Q := \gamma P(\Pi_Q - \Pi_{Q^*})Q^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}.$$

Then Q-VI can be represented as

$$Q_{k+1} - Q^* = A_{Q_k}(Q_k - Q^*) + b_{Q_k}, \tag{6}$$

where A_{Q_k} and b_{Q_k} switch among matrices from $\{\gamma P \Pi^\pi : \pi \in \Theta\}$ and vectors from $\{\gamma P(\Pi^\pi - \Pi^{\pi^*})Q^* : \pi \in \Theta\}$ according to the changes of Q_k.

To make this switching description finite-dimensional in the usual switching-system notation, let $\varphi : \Theta \to \{1, 2, \ldots, |\Theta|\}$ be a one-to-one mapping from a deterministic policy $\pi \in \Theta$ to an integer in $\{1, 2, \ldots, |\Theta|\}$. For $i = \varphi(\pi)$, define

$$A_i = \gamma P \Pi^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}, \qquad b_i = \gamma P(\Pi^\pi - \Pi^{\pi^*})Q^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}.$$

Then Equation (6) is an affine switching system with switching signal $\sigma_k \in \{1, 2, \ldots, |\Theta|\}$ determined by $\sigma_k = \varphi(\pi_k)$, where

$$\pi_k(\cdot) := \arg\max_{a \in \mathcal{A}} Q_k(\cdot, a) \in \Theta.$$

Two elementary facts connect this affine switching model with the classical contraction rate. The first gives a uniform norm bound for every active subsystem, and the second identifies the JSR of the full switching family.

Lemma 2 (Lee et al. (2023)). *For any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\|A_Q\|_\infty = \gamma$, where the matrix norm $\|A\|_\infty := \max_i \sum_j |A_{ij}|$ and A_{ij} is the element of A in the i-th row and j-th column.*

Proof. Note that $\sum_j |[A_Q]_{ij}| = \sum_j |[\gamma P \Pi_Q]_{ij}| = \gamma$, which completes the proof. □

This bound also identifies the exact JSR of the linear part of the full switching family.

Lemma 3. *The JSR of the linear part $\{A_i\}_{i=1}^M$ of the affine switching representation Equation (6) is γ.*

Proof. Since each $P\Pi^\pi$ is row-stochastic, for every switching sequence $\bar{\sigma}_k = (\sigma_1, \ldots, \sigma_k)$, the matrix

$$A_{\sigma_k} \cdots A_{\sigma_1} = \gamma^k (P\Pi^{\pi_{\sigma_k}}) \cdots (P\Pi^{\pi_{\sigma_1}})$$

has infinity norm equal to γ^k. Hence,

$$\rho(A_1, \cdots, A_M) = \lim_{k\to\infty} \max_{\bar{\sigma}_k \in \{1,\ldots,M\}^k} \|A_{\sigma_k} \cdots A_{\sigma_1}\|_\infty^{1/k} = \lim_{k\to\infty} (\gamma^k)^{1/k} = \gamma.$$

This completes the proof. □

B Structural conditions for $\bar{\rho} < \gamma$

The preceding analysis established that the restricted JSR always satisfies $\bar{\rho} \le \gamma$. We now give structural conditions under which the inequality becomes strict. These conditions are not imposed in the main results, but they clarify when the projected state-action dynamics possess uniform mixing in directions transverse to $\mathrm{span}(\mathbf{1})$. The analysis is based on standard coefficients of ergodicity, scrambling matrix arguments, and span-seminorm contractions for products of stochastic matrices Dobrushin (1956); Hajnal & Bartlett (1958); Wolfowitz (1963); Seneta (2006).

One subtle point is that the exact Q-VI error representation in Lemma 3 is written in terms of stochastic policies. Nevertheless, the strict inequality $\bar{\rho} < \gamma$ can be characterized using deterministic policy products. The reason is that the stochastic-policy matrices form the convex hull of the deterministic-policy matrices, and the JSR is invariant under convexification Rota & Strang (1960); Blondel & Nesterov (2005).

We denote the state-action index set in this section by

$$\mathcal{Y} := \mathcal{S} \times \mathcal{A}.$$

For each deterministic policy $\pi \in \Theta$, define

$$B_\pi := P\Pi^\pi,$$

so that $A_\pi = \gamma B_\pi$. Then, B_π is row-stochastic on the finite index set $\mathcal{Y}$. For each stochastic policy $\mu : \mathcal{S} \to \Delta_{|\mathcal{A}|}$, define similarly

$$B_\mu := P\Pi^\mu.$$

Then, B_μ is also row-stochastic. Moreover, every stochastic-policy matrix is a convex combination of deterministic-policy matrices:

$$B_\mu = \sum_{\pi\in\Theta} c_\pi(\mu) B_\pi, \qquad c_\pi(\mu) := \prod_{s\in\mathcal{S}} \mu(\pi(s)|s),$$

where $c_\pi(\mu) \ge 0$ and $\sum_{\pi\in\Theta} c_\pi(\mu) = 1$. This proves

$$\{B_\mu : \mu : \mathcal{S} \to \Delta_{|\mathcal{A}|}\} \subseteq \mathrm{co}\{B_\pi : \pi \in \Theta\}.$$

The reverse inclusion also holds. Indeed, let $\sum_{\pi\in\Theta} c_\pi B_\pi$ be an arbitrary convex combination, with $c_\pi \ge 0$ and $\sum_{\pi\in\Theta} c_\pi = 1$. Define a stochastic policy $\mu_c : \mathcal{S} \to \Delta_{|\mathcal{A}|}$ by

$$\mu_c(a|s) := \sum_{\pi\in\Theta:\, \pi(s)=a} c_\pi, \qquad s \in \mathcal{S}, \quad a \in \mathcal{A}.$$

Then, row by row in the definition of the action transition matrix,

$$\Pi^{\mu_c} = \sum_{\pi \in \Theta} c_\pi \Pi^\pi.$$

Consequently,

$$B_{\mu_c} = P\Pi^{\mu_c} = \sum_{\pi \in \Theta} c_\pi P\Pi^\pi = \sum_{\pi \in \Theta} c_\pi B_\pi.$$

Therefore,

$$\{B_\mu : \mu : \mathcal{S} \to \Delta_{|\mathcal{A}|}\} = \mathrm{co}\{B_\pi : \pi \in \Theta\}.$$

Let

$$\widehat{B}_\pi := \mathbf{\Pi}_\perp B_\pi \mathbf{\Pi}_\perp, \qquad \bar{B}_\mu := \mathbf{\Pi}_\perp B_\mu \mathbf{\Pi}_\perp.$$

Then, it follows that

$$\bar{B}_\mu = \sum_{\pi \in \Theta} c_\pi(\mu) \bar{B}_\pi,$$

and by the convex-hull invariance of the JSR,

$$\rho(\bar{B}_\mu : \mu : \mathcal{S} \to \Delta_{|\mathcal{A}|}) = \rho(\bar{B}_\pi : \pi \in \Theta).$$

Since

$$\bar{A}_\pi = \mathbf{\Pi}_\perp A_\pi \mathbf{\Pi}_\perp = \gamma \mathbf{\Pi}_\perp B_\pi \mathbf{\Pi}_\perp = \gamma \bar{B}_\pi,$$

we have

$$\bar{\rho} = \rho(\bar{A}_\pi : \pi \in \Theta) = \gamma \rho(\bar{B}_\pi : \pi \in \Theta).$$

For a deterministic policy sequence

$$\omega = (\pi_0, \pi_1, \ldots, \pi_{L-1}) \in \Theta^L,$$

define the corresponding product

$$B_\omega := B_{\pi_{L-1}} \cdots B_{\pi_1} B_{\pi_0}.$$

Equivalently, the directed graph of B_π has an edge

$$(s, a) \longrightarrow_\pi (s', a')$$

if and only if $P(s'|s, a) > 0$ and $a' = \pi(s')$. For a policy sequence ω and $x \in \mathcal{Y}$, let

$$\mathcal{R}_\omega(x) := \{y \in \mathcal{Y} : (B_\omega)_{xy} > 0\}$$

be the set of state-action nodes reachable from x by following the policies $\pi_{L-1}, \ldots, \pi_0$ in that order. This matches the product convention for the column-vector error recursion; forward transitions under ω instead correspond to $B_{\pi_0} \cdots B_{\pi_{L-1}}$.

Definition 2 (Scrambling matrix Hajnal & Bartlett (1958); Wolfowitz (1963); Seneta (2006))**.** *A row-stochastic matrix $B \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|}$ is called scrambling if every pair of rows has at least one common positive column. Equivalently, for every $x, y \in \mathcal{Y}$, there exists $v \in \mathcal{Y}$ such that*

$$B_{xv} > 0 \qquad \textit{and} \qquad B_{yv} > 0.$$

For a product B_ω, this is equivalent to

$$\mathcal{R}_\omega(x) \cap \mathcal{R}_\omega(y) \neq \varnothing, \qquad \forall x, y \in \mathcal{Y}.$$

Definition 3 (Dobrushin coefficient Dobrushin (1956); Gaubert & Qu (2015))**.** *For a row-stochastic matrix $B \in \mathbb{R}^{|\mathcal{Y}| \times |\mathcal{Y}|}$, its Dobrushin coefficient is defined by*

$$\tau_{\mathrm{D}}(B) := \frac{1}{2} \max_{x,y \in \mathcal{Y}} \sum_{v \in \mathcal{Y}} |B_{xv} - B_{yv}|.$$

Equivalently,

$$\tau_{\mathrm{D}}(B) = 1 - \min_{x,y \in \mathcal{Y}} \sum_{v \in \mathcal{Y}} \min\{B_{xv}, B_{yv}\}.$$

Definition 4 (Diameter seminorm Gaubert & Qu (2015); Tsitsiklis et al. (1986))**.** *For a vector $v \in \mathbb{R}^{|\mathcal{Y}|}$, the diameter seminorm is defined as*

$$\|v\|_{\mathrm{dm}} := \max_{x \in \mathcal{Y}} v_x - \min_{x \in \mathcal{Y}} v_x.$$

It is a seminorm on $\mathbb{R}^{|\mathcal{Y}|}$, and it becomes a norm on the quotient space $\mathbb{R}^{|\mathcal{Y}|}/\operatorname{span}(\mathbf{1})$, since $\|v\|_{\mathrm{dm}} = 0$ if and only if $v \in \operatorname{span}(\mathbf{1})$.

We use the following standard properties of the Dobrushin coefficient and the diameter seminorm.

Lemma 14 (Dobrushin coefficient and diameter-seminorm contraction Dobrushin (1956); Hajnal & Bartlett (1958); Seneta (2006))**.** *Let B and C be row-stochastic matrices of size $|\mathcal{Y}| \times |\mathcal{Y}|$. Then the following statements hold:*

1. $0 \le \tau_{\mathrm{D}}(B) \le 1$.
2. $\tau_{\mathrm{D}}(B) < 1$ *if and only if B is scrambling.*
3. $\tau_{\mathrm{D}}(BC) \le \tau_{\mathrm{D}}(B)\tau_{\mathrm{D}}(C)$.
4. *If B_i are row-stochastic and $\alpha_i \ge 0$, $\sum_i \alpha_i = 1$, then*

$$\tau_{\mathrm{D}}\left(\sum_i \alpha_i B_i\right) \le \sum_i \alpha_i \tau_{\mathrm{D}}(B_i).$$

   In particular, τ_{D} is convex on the set of row-stochastic matrices.

5. *The Dobrushin coefficient is the induced operator norm of B on the quotient space $\mathbb{R}^{|\mathcal{Y}|}/\operatorname{span}(\mathbf{1})$ equipped with the diameter seminorm:*

$$\tau_{\mathrm{D}}(B) = \sup_{v \notin \operatorname{span}(\mathbf{1})} \frac{\|Bv\|_{\mathrm{dm}}}{\|v\|_{\mathrm{dm}}}.$$

   Consequently, for every $v \in \mathbb{R}^{|\mathcal{Y}|}$,

$$\|Bv\|_{\mathrm{dm}} \le \tau_{\mathrm{D}}(B)\|v\|_{\mathrm{dm}}.$$

6. *For every $v \in \mathbb{R}^{|\mathcal{Y}|}$ and every $c \in \mathbb{R}$,*

$$\|v + c\mathbf{1}\|_{\mathrm{dm}} = \|v\|_{\mathrm{dm}}.$$

   Consequently, if $\mathbf{\Pi}_\perp = I - \frac{1}{|\mathcal{Y}|}\mathbf{1}\mathbf{1}^\top$, then

$$\sup_{v \notin \operatorname{span}(\mathbf{1})} \frac{\|\mathbf{\Pi}_\perp B \mathbf{\Pi}_\perp v\|_{\mathrm{dm}}}{\|v\|_{\mathrm{dm}}} = \tau_{\mathrm{D}}(B).$$

Proof. See Appendix D.18.

We can now state an if-and-only-if characterization of strictness for the full restricted switching family.

Proposition 3 (Characterization of $\bar{\rho} < \gamma$ by uniform scrambling)**.** *The following statements are equivalent:*

1. $\bar{\rho} < \gamma$.
2. *There exists an integer $L \ge 1$ such that, for every deterministic policy sequence $\omega = (\pi_0, \dots, \pi_{L-1}) \in \Theta^L$, the product B_ω is scrambling.*
3. *There exists an integer $L \ge 1$ such that*

$$\max_{\omega \in \Theta^L} \tau_{\mathrm{D}}(B_\omega) < 1.$$

4. *Equivalently, there exists an integer $L \geq 1$ such that*

$$\sup_{\mu_0,\dots,\mu_{L-1}} \tau_{\mathrm{D}}\left(B_{\mu_{L-1}} \cdots B_{\mu_1} B_{\mu_0}\right) < 1,$$

where the supremum is taken over all stochastic policies $\mu_0, \dots, \mu_{L-1} : \mathcal{S} \to \Delta_{|\mathcal{A}|}$.

Moreover, if these equivalent conditions hold and

$$\eta_L := \min_{\omega \in \Theta^L} \min_{x,y \in \mathcal{Y}} \sum_{v \in \mathcal{Y}} \min\{(B_\omega)_{xv}, (B_\omega)_{yv}\},$$

then $\eta_L > 0$ and

$$\bar{\rho} \leq \gamma(1 - \eta_L)^{1/L} < \gamma.$$

Proof. See Appendix D.19.

Proposition 3 shows that the bounded scrambling condition is a necessary and sufficient characterization of $\bar{\rho} < \gamma$ for the full restricted switching family. It can be stronger than what is necessary for a particular Q-VI trajectory, since not every admissible stochastic switching sequence must be realized by the Bellman maximization dynamics along a given initialization.

A simple sufficient condition for the equivalent properties in Proposition 3 is a one-step common descendant. Suppose that there exists a state $s' \in \mathcal{S}$ such that

$$p_{\min} := \min_{s \in \mathcal{S},\ a \in \mathcal{A}} P(s'|s,a) > 0.$$

Then, for every deterministic policy π, all rows of B_π assign at least $p_{\min}$ mass to the common state-action column $(s', \pi(s'))$. Hence, B_π is one-step scrambling and

$$\bar{\rho} \leq \gamma(1 - p_{\min}) < \gamma.$$

More generally, if there exist $\varepsilon_{\mathrm{D}} > 0$ and a probability distribution $\nu \in \Delta_{|\mathcal{S}|}$ such that the Doeblin-type minorization condition

$$P(\cdot|s,a) \geq \varepsilon_{\mathrm{D}} \nu(\cdot), \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

holds componentwise, then every B_π has a common stochastic component of mass at least ε_{D}. More precisely, for each fixed deterministic policy π, each row of B_π dominates the same subprobability measure on $\mathcal{Y}$ that assigns mass $\varepsilon_{\mathrm{D}} \nu(s')$ to $(s', \pi(s'))$. This is a standard uniform mixing condition for finite Markov chains Seneta (2006). Consequently,

$$\bar{\rho} \leq \gamma(1 - \varepsilon_{\mathrm{D}}) < \gamma.$$

These one-step conditions are strong but easy to check. They are sufficient, not necessary: $\bar{\rho} < \gamma$ may still hold when row overlap appears only after several steps, as captured by the bounded scrambling condition in Proposition 3.

The same characterization also shows what can prevent strictness. For a fixed deterministic policy π, define the induced state transition matrix

$$P_\pi := \Pi^\pi P \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}.$$

The nonzero eigenvalues of $B_\pi = P\Pi^\pi$ coincide with those of P_π. Consequently, if some deterministic policy π has $|\lambda_2(P_\pi)| = 1$, then, by Lemma 9, $\rho(\bar{A}_\pi) = \gamma$. Since $\bar{A}_\pi$ is one member of the full restricted family, it follows that $\bar{\rho} \geq \rho(\bar{A}_\pi) = \gamma$. Together with Lemma 5, this gives $\bar{\rho} = \gamma$. Therefore, strict inequality for the full family is impossible if even one deterministic policy induces a state Markov chain with a nontrivial unit-modulus mode, where a nontrivial unit-modulus mode means an eigenvalue λ of P_π with $|\lambda| = 1$, other than the Perron eigenvalue associated with the invariant constant direction. Such a mode prevents contraction on the quotient space obtained after removing $\mathrm{span}(\mathbf{1})$ Seneta (2006); Levin & Peres (2017).

Standard Markov-chain theory shows that this obstruction occurs, for example, when P_π has more than one closed communicating class, meaning a communicating class with no transition to states outside the class, or when a closed communicating class is periodic, meaning that the greatest common divisor of its possible return times is larger than one Seneta (2006); Levin & Peres (2017).

Therefore, a necessary structural requirement for the full inequality $\bar{\rho} < \gamma$ is that every deterministic stationary policy be aperiodic and unichain, where unichain means that there is exactly one closed communicating class, possibly together with transient states. This single-policy requirement alone, however, is not sufficient for arbitrary switching; the precise full-family condition is the bounded scrambling condition in Proposition 3.

This full-family condition should be distinguished from the post-identification condition in Theorem 2. For all $k \geq K_{\mathrm{id}}$, only the optimal restricted family $\mathcal{H}_* = \{\bar{A}_\pi : \pi \in \Theta^*\}$ remains active. Hence, one only needs bounded scrambling over optimal policies to obtain $\bar{\rho}_* < \gamma$. In the unique-optimal-policy case, this reduces to the single-chain condition

$$|\lambda_2(P_{\pi^*})| < 1,$$

or equivalently, in standard finite-state Markov-chain terminology, that the optimal-policy chain has one aperiodic closed communicating class, possibly with transient states. This recovers the spectral statement in Corollary 3 and gives a graph-theoretic interpretation of when the post-identification transverse decay is strictly faster than the classical γ-rate.

C Examples

This section illustrates the geometric picture developed above using a small discounted MDP with state set $\mathcal{S} = \{1, 2, 3\}$, action set $\mathcal{A} = \{1, 2\}$, and discount factor $\gamma = 0.95$. The transition probability matrices and the expected one-step rewards are chosen as

$$P_1 = \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.3 & 0.6 \end{bmatrix}, \qquad P_2 = \begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.4 & 0.3 & 0.3 \\ 0.3 & 0.3 & 0.4 \end{bmatrix}, \qquad R = \begin{bmatrix} 1.0 & 0.2 \\ 0.6 & 0.0 \\ 1.2 & 0.3 \end{bmatrix},$$

respectively. Here $P_a = P(\cdot \mid \cdot, a)$ for $a \in \mathcal{A} = \{1, 2\}$, and the (s, a)-entry of R represents the expected reward at state $s \in \mathcal{S} = \{1, 2, 3\}$ under action $a \in \mathcal{A}$, i.e. $R(s, a)$. For this example, the unique optimal deterministic policy is

$$\pi^* = \big(\pi^*(1), \pi^*(2), \pi^*(3)\big) = (1, 1, 1),$$

and the corresponding optimal Q-function is

$$Q^* = \begin{bmatrix} 18.2229 & 17.3026 \\ 17.6194 & 17.2172 \\ 18.4947 & 17.5430 \end{bmatrix},$$

where the (s, a)-entry represents the optimal Q-function value at state s under action a, i.e. $Q^*(s, a)$. The corresponding minimum optimality gap is

$$\Delta := \min_{s \in \mathcal{S}_{\mathrm{sep}}} \bar{\Delta}_s, \qquad \bar{\Delta}_s := V^*(s) - \max_{a \notin \Phi^*(s)} Q^*(s, a) \approx 0.4022.$$

We consider the affine space $\mathcal{X}_1 = Q^* + \mathrm{span}(\mathbf{1})$, and choose the tube radius $\delta = 0.4\Delta \approx 0.1609$, which satisfies $\delta < \Delta/2$. To compare the convergence to the optimal Q-function and the convergence to the affine set $\mathcal{X}_1$, we plot the normalized quantities

$$\frac{\|Q_k - Q^*\|_\infty}{\|Q_0 - Q^*\|_\infty} \qquad \text{and} \qquad \frac{\mathrm{dist}_2(Q_k, \mathcal{X}_1)}{\mathrm{dist}_2(Q_0, \mathcal{X}_1)}.$$

As reference curves, we also include γ^k and $(\gamma|\lambda_2|)^k$, where $|\lambda_2|$ denotes the modulus of the second largest eigenvalue of the state-action transition matrix $P\Pi_{\pi^*}$ associated with the optimal policy. For this example, $|\lambda_2| \approx 0.5618$ and $\gamma|\lambda_2| \approx 0.5337$.

To visualize the geometry in two dimensions, we consider the two-dimensional affine plane

$$Q^* + \mathrm{span}\{\hat{\mathbf{1}}, \hat{d}\},$$

where

$$\hat{\mathbf{1}} := \frac{1}{\sqrt{6}}(1, 1, 1, 1, 1, 1)^\top \in \mathbb{R}^6, \qquad \hat{d} := \frac{1}{\sqrt{2}}(1, 0, 0, -1, 0, 0)^\top \in \mathbb{R}^6.$$

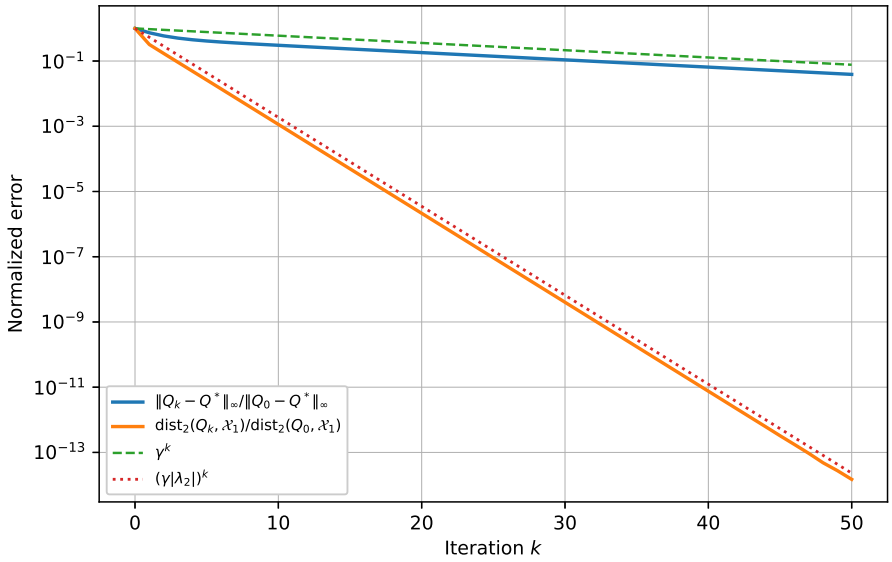


Figure 2: Normalized comparison of $\|Q_k - Q^*\|_\infty$ and $\mathrm{dist}_2(Q_k, \mathcal{X}_1)$ for the toy MDP. The dashed curves indicate the reference rates γ^k and $(\gamma|\lambda_2|)^k$. The plot shows that the iterate approaches the affine set $\mathcal{X}_1$ faster than it approaches the optimal Q-function Q^*.

Here, $\hat{\mathbf{1}}$ is the normalized all-ones direction, which is tangent to $\mathcal{X}_1$, while $\hat{d}$ is a transverse direction that perturbs the action contrast at the first state. For any Q, write the coordinates of $Q - Q^*$ on this plane as

$$u(Q) := \langle Q - Q^*, \hat{\mathbf{1}} \rangle, \qquad v(Q) := \langle Q - Q^*, \hat{d} \rangle.$$

For the single-trajectory experiment, the initial condition is chosen on this tilted plane as follows:

$$Q_0 = \begin{bmatrix} 19.5495 & 16.6292 \\ 17.9460 & 17.5438 \\ 18.8213 & 17.8696 \end{bmatrix}.$$

For this particular Q_0, the tie-broken greedy action is already the optimal action at every state. Thus, this single trajectory is used to visualize finite-time entrance into the invariant tube, rather than to demonstrate first-time policy identification from a non-optimal greedy policy. Starting from this Q_0, we run Q-VI for 50 iterations. Figure 2 shows the normalized decay of $\|Q_k - Q^*\|_\infty$ and $\mathrm{dist}_2(Q_k, \mathcal{X}_1)$. The plot shows that the distance to $\mathcal{X}_1$ decreases more rapidly than the full Q-function error in the transient regime, in line with the theoretical picture developed in this paper. Figure 3 shows the orthogonal projection of the full Q-VI trajectory onto the tilted plane. The dashed line represents the slice of $\mathcal{X}_1$ and the shaded strip is the slice of $\mathcal{T}_\delta$. The projected trajectory starts from the initial point, enters the strip in finite time, and then converges to the origin, which corresponds to $Q_k \to Q^*$.

To show the global geometry more clearly, we also consider 12 initial conditions placed uniformly on the circle of radius 2 in the tilted plane. For each such initial condition, we run Q-VI for 50 iterations and project the resulting trajectory onto the same tilted plane. Figure 4 shows these projected trajectories together with the slice of $\mathcal{X}_1$, the slice of $\mathcal{T}_\delta$, and the initial circle. This figure provides a global view of trajectories from different directions being rapidly drawn toward the affine set $\mathcal{X}_1$ before ultimately converging to Q^*.

We next present a tabular Q-learning example on the same toy MDP used in the preceding Q-VI example. The underlying discounted MDP, the optimal policy $\pi^* = \big(\pi^*(1), \pi^*(2), \pi^*(3)\big) = (1, 1, 1)$, the optimal Q-function Q^*, the affine set $\mathcal{X}_1$, and the tube radius δ are the same as in the preceding example, so we do not repeat their definitions. We consider the standard asynchronous tabular Q-learning update. In the simulation, the action is sampled from the uniform behavior policy $\mu(a \mid s) = \frac{1}{|\mathcal{A}|}$ for $s \in \mathcal{S}$ and $a \in \mathcal{A}$, and the step size is chosen as $\alpha_t = \frac{0.35}{1+0.01\,t}$. To visualize

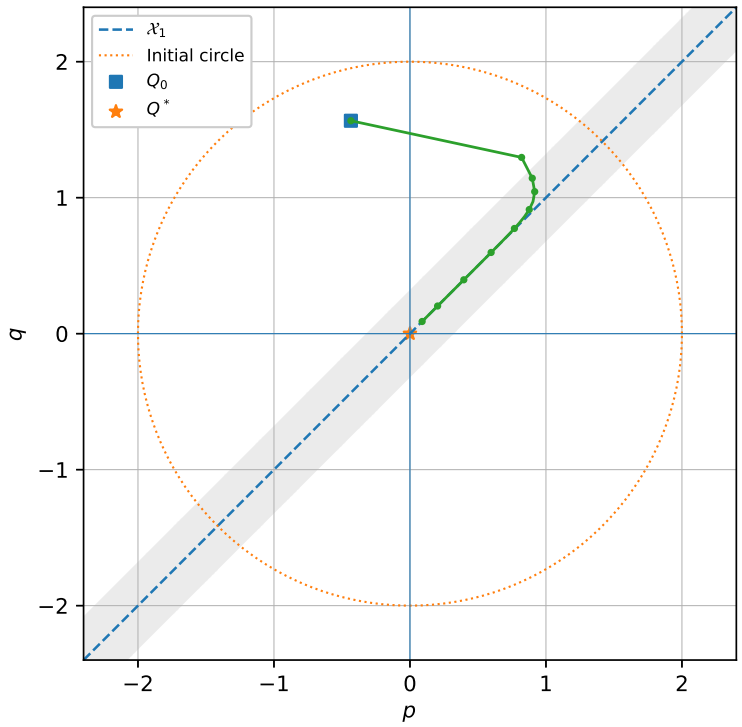


Figure 3: Orthogonal projection of a single full Q-VI trajectory onto the tilted plane. The dashed line is the slice of $\mathcal{X}_1$ and the shaded strip is the slice of the tube $\mathcal{T}_\delta$. The projected trajectory starts from the initial point, enters the strip, and then converges to the origin.

the geometry, we consider the two-dimensional affine plane identical to the previous Q-VI case. As before, we choose 12 initial conditions uniformly on the boundary of a circle of radius $r = 2$ in the tilted plane.

Figure 5 shows the projected sample paths corresponding to these 12 initial conditions, together with the line $q = p$, which is the projection of $\mathcal{X}_1$, the strip corresponding to the projected tube $\mathcal{T}_\delta$, and the boundary of the initial circle. Since Q-learning is stochastic, the trajectories are not monotone and exhibit path-dependent fluctuations. Nevertheless, a clear common trend can be observed: from a variety of initial directions, the trajectories are first drawn toward the line $q = p$, enter the projected tube, and then continue to move toward the origin. These paths are consistent with the geometric interpretation from the Q-VI example: even in the stochastic tabular Q-learning setting, the affine set $\mathcal{X}_1$ acts as a transient attracting set before the iterates refine their convergence toward Q^*.

Compared with the Q-VI example, the present figure shows a family of stochastic sample paths rather than a single deterministic trajectory. This makes the same geometric mechanism easier to see: although the noise causes trajectory-dependent oscillations, the iterates tend to approach the neighborhood of $\mathcal{X}_1$ relatively quickly and only afterward progress toward the optimal Q-function Q^*.

D Auxiliary Statements and Proofs

D.1 Statement and Proof of Lemma 4

Lemma 4. *We have* $\mathcal{X}_1 \subset \mathcal{X}^*$.

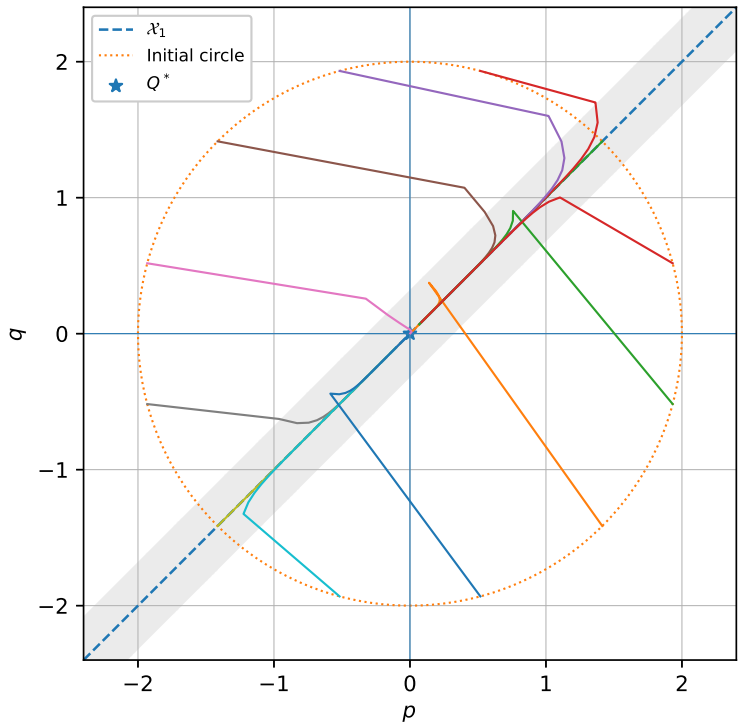


Figure 4: Orthogonal projections of 12 full Q-VI trajectories onto the tilted plane, where the initial conditions are chosen uniformly on a circle of radius 2 in the plane. The dashed line is the slice of $\mathcal{X}_1$, the shaded strip is the slice of $\mathcal{T}_\delta$, and the dotted curve is the initial circle. The plot gives a global geometric view of trajectories being first attracted toward $\mathcal{X}_1$ and then converging to Q^*.

Proof. Let $Q \in \mathcal{X}_1$. Then $Q = Q^* + \alpha\mathbf{1}$ for some $\alpha \in \mathbb{R}$. Since adding a constant multiple of $\mathbf{1}$ does not change the action ordering,

$$\operatorname{Arg\,max}_{a\in\mathcal{A}} Q(s,a) = \operatorname{Arg\,max}_{a\in\mathcal{A}} Q^*(s,a) = \Phi^*(s), \qquad \forall s \in \mathcal{S}.$$

Hence every tie-broken greedy action of Q belongs to $\Phi^*(s)$, so $Q \in \mathcal{X}^*$. This completes the proof. □

D.2 Statement and Proof of Proposition 1

Proposition 1 (Invariance of $\mathcal{X}_1$)**.** *If $Q_k \in \mathcal{X}_1$, then $Q_{k+1} \in \mathcal{X}_1$. More precisely, if $Q_k = Q^* + \alpha_k\mathbf{1}$ for some $\alpha_k \in \mathbb{R}$, then $Q_{k+1} = Q^* + \gamma\alpha_k\mathbf{1}$.*

Proof. Suppose that $Q_k = Q^* + \alpha_k\mathbf{1}$. Since adding a constant multiple of $\mathbf{1}$ does not change the action ordering, every tie-broken greedy action of Q_k is optimal for Q^*. Hence $\pi_{Q_k} \in \Theta^*$. Using the Bellman update, $Q_{k+1} = R + \gamma P\Pi_{Q_k}(Q^* + \alpha_k\mathbf{1})$. Because π_{Q_k} is an optimal policy, we have $R + \gamma P\Pi_{Q_k}Q^* = Q^*$. In addition, since $P\Pi_{Q_k}$ is stochastic, it satisfies $P\Pi_{Q_k}\mathbf{1} = \mathbf{1}$. Therefore, it follows that $Q_{k+1} = Q^* + \gamma\alpha_k\mathbf{1}$, and $Q_{k+1} \in \mathcal{X}_1$, which proves the claim. □

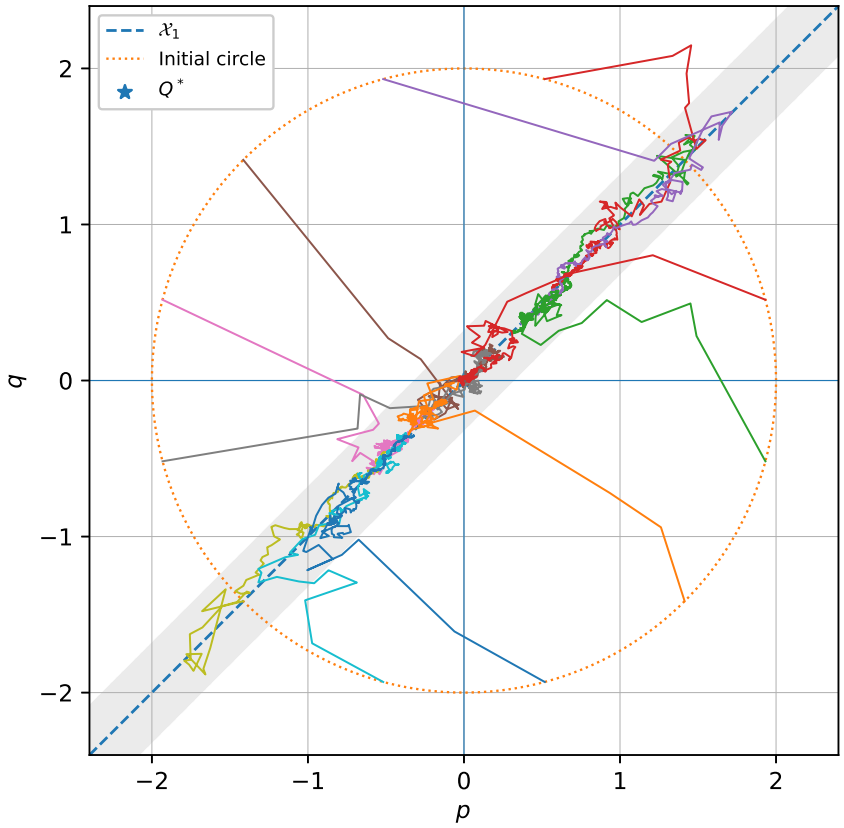


Figure 5: Two-dimensional projection of tabular Q-learning trajectories onto the tilted plane $Q^* + \mathrm{span}\{\hat{\mathbf{1}}, \hat{d}\}$, shown in the rotated coordinates $p = (u - v)/\sqrt{2}$ and $q = (u + v)/\sqrt{2}$. The dashed line $q = p$ is the projection of $\mathcal{X}_1$, and the shaded strip is the projection of the tube $\mathcal{T}_\delta$, namely $|q - p| \le 2\delta$. The dotted circle indicates the boundary of the common initial set in the (u, v)-plane. Starting from 12 different initial points on this circle, the stochastic Q-learning trajectories tend to enter the strip and then move toward the origin, which corresponds to Q^*.

D.3 Proof of Proposition 2

Proof. Let $Q \in \mathcal{T}_\delta$. By definition, there exists $\alpha \in \mathbb{R}$ such that $\|Q - (Q^* + \alpha\mathbf{1})\|_\infty \le \delta$. Fix any state $s \in \mathcal{S}_{\mathrm{sep}}$ and any action $b \notin \Phi^*(s)$. Then, we can derive the following inequalities:

$$\begin{aligned}
\max_{a \in \Phi^*(s)} Q(s,a) - Q(s,b) =& \max_{a \in \mathcal{A}} Q^*(s,a) - Q^*(s,b) \\
&+ \left(\max\nolimits_{a \in \Phi^*(s)} Q(s,a) - (\max\nolimits_{a \in \mathcal{A}} Q^*(s,a) + \alpha)\right) \\
&- (Q(s,b) - (Q^*(s,b) + \alpha)) \\
\ge& \max_{a \in \mathcal{A}} Q^*(s,a) - Q^*(s,b) \\
&+ \left(\max\nolimits_{a \in \Phi^*(s)} Q(s,a) - \left(\max\nolimits_{a \in \Phi^*(s)} Q^*(s,a) + \alpha\right)\right) \\
&- (Q(s,b) - (Q^*(s,b) + \alpha)) \\
\ge& \bar{\Delta}_s - 2\delta \\
\ge& \bar{\Delta} - 2\delta \\
>& 0.
\end{aligned}$$

Therefore no non-optimal action can be tie-broken greedy at any state $s \in \mathcal{S}_{\mathrm{sep}}$. For states outside $\mathcal{S}_{\mathrm{sep}}$, all actions are optimal by definition. Therefore, $\pi_Q(s) \in \Phi^*(s)$ for all $s \in \mathcal{S}$, and $Q \in \mathcal{X}^*$. Now let $Y := Q^* + \alpha\mathbf{1} \in \mathcal{X}_1$ with $\|Q - Y\|_\infty \le \delta$. Since $Q \in \mathcal{X}^*$, we have $\pi_Q \in \Theta^*$. Although

π_Q need not be the tie-broken greedy policy of Y, the optimality of π_Q implies

$$R + \gamma P \Pi_Q Y = R + \gamma P \Pi_Q (Q^* + \alpha \mathbf{1}) = Q^* + \gamma \alpha \mathbf{1} \in \mathcal{X}_1.$$

Moreover,

$$F(Q) - \big(R + \gamma P \Pi_Q Y\big) = \gamma P \Pi_Q (Q - Y).$$

Taking the infinity norm and using $\|P\Pi_Q\|_\infty = 1$, we obtain

$$\mathrm{dist}_\infty(F(Q), \mathcal{X}_1) \le \big\|F(Q) - \big(R + \gamma P \Pi_Q Y\big)\big\|_\infty \le \gamma \|Q - Y\|_\infty \le \gamma\delta < \delta.$$

Therefore, one concludes that $F(\mathcal{T}_\delta) \subset \mathcal{T}_{\gamma\delta} \subset \mathcal{T}_\delta$. This completes the proof. □

D.4 Statement and Proof of Corollary 1

Corollary 1 (Basic finite-time entrance into $\mathcal{X}^*$)**.** *Let Assumption 1 hold. Define*

$$K_{\mathrm{basic}} := \begin{cases} 0, & \|Q_0 - Q^*\|_\infty < \dfrac{\bar{\Delta}}{2}, \\ \left\lfloor \dfrac{\log\left(\dfrac{2\|Q_0 - Q^*\|_\infty}{\bar{\Delta}}\right)}{-\log\gamma} \right\rfloor + 1, & \|Q_0 - Q^*\|_\infty \ge \dfrac{\bar{\Delta}}{2}. \end{cases}$$

Then, we have $Q_k \in \mathcal{X}^*$ for all $k \ge K_{\mathrm{basic}}$.

Proof. By Equation (2), one has $\|Q_k - Q^*\|_\infty \le \gamma^k \|Q_0 - Q^*\|_\infty$. The definition of K_{basic} guarantees

$$\gamma^k \|Q_0 - Q^*\|_\infty < \frac{\bar{\Delta}}{2}, \qquad \forall k \ge K_{\mathrm{basic}}.$$

Hence, for all $k \ge K_{\mathrm{basic}}$, we get $\|Q_k - Q^*\|_\infty < \frac{\bar{\Delta}}{2}$. Since $Q^* \in \mathcal{X}_1$, this implies $\mathrm{dist}_\infty(Q_k, \mathcal{X}_1) < \frac{\bar{\Delta}}{2}$. For each such k, choose a number δ_k satisfying

$$\mathrm{dist}_\infty(Q_k, \mathcal{X}_1) \le \delta_k < \frac{\bar{\Delta}}{2},$$

which is possible because the distance is strictly smaller than $\bar{\Delta}/2$. Then $Q_k \in \mathcal{T}_{\delta_k}$, and by Proposition 2 we have $\mathcal{T}_{\delta_k} \subset \mathcal{X}^*$. Hence $Q_k \in \mathcal{X}^*$ for all $k \ge K_{\mathrm{basic}}$, which completes the proof. □

D.5 Proof of Lemma 3

Proof. For each state $s \in \mathcal{S}$, define the optimality error

$$\delta_k(s) := \max_{a \in \mathcal{A}} Q_k(s,a) - \max_{a \in \mathcal{A}} Q^*(s,a).$$

Since $Q_k(s,a) = Q^*(s,a) + e_k(s,a)$, we have

$$\min_{a \in \mathcal{A}} e_k(s,a) \le \delta_k(s) \le \max_{a \in \mathcal{A}} e_k(s,a).$$

Therefore, $\delta_k(s)$ belongs to the convex hull of the finite set $\{e_k(s,a) : a \in \mathcal{A}\}$. Hence there exists a probability vector $\mu_k(\cdot|s) \in \Delta_{|\mathcal{A}|}$ for each $k \ge 0$ such that

$$\delta_k(s) = \sum_{a \in \mathcal{A}} \mu_k(a|s) e_k(s,a) = \mu_k(\cdot|s)^\top e_k(s,\cdot).$$

Stacking these equalities over all states and using the definition of Π^{μ_k} in Equation (3), with the same action-major ordering as the vector e_k, yields

$$\delta_k = \begin{bmatrix} \delta_k(1) \\ \delta_k(2) \\ \vdots \\ \delta_k(|\mathcal{S}|) \end{bmatrix} = \begin{bmatrix} \mu_k(\cdot \mid 1)^\top \otimes e_1^\top \\ \mu_k(\cdot \mid 2)^\top \otimes e_2^\top \\ \vdots \\ \mu_k(\cdot \mid |\mathcal{S}|)^\top \otimes e_{|\mathcal{S}|}^\top \end{bmatrix} \begin{bmatrix} e_k(\cdot,1) \\ e_k(\cdot,2) \\ \vdots \\ e_k(\cdot,|\mathcal{A}|) \end{bmatrix} = \Pi^{\mu_k} e_k \in \mathbb{R}^{|\mathcal{S}|}.$$

Now subtract the Bellman optimality equation from the Q-VI update:

$$\begin{aligned} e_{k+1}(s,a) &= \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a) \left(\max_{b \in \mathcal{A}} Q_k(s',b) - \max_{b \in \mathcal{A}} Q^*(s',b) \right) \\ &= \gamma \sum_{s' \in \mathcal{S}} P(s'|s,a)\, \delta_k(s'). \end{aligned}$$

In vector form, we have

$$e_{k+1} = \gamma P \delta_k = \gamma P \Pi^{\mu_k} e_k = A_{\mu_k} e_k.$$

This proves the claim. □

D.6 Proof of Lemma 4

Proof. Since $z_k := \Pi_\perp e_k$, we have the orthogonal decomposition

$$e_k = (I - \Pi_\perp) e_k + \Pi_\perp e_k.$$

Using $\Pi_\perp = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$, it follows that

$$(I - \Pi_\perp) e_k = \frac{1}{n}\mathbf{1}\mathbf{1}^\top e_k = \left(\frac{1}{n}\mathbf{1}^\top e_k \right) \mathbf{1}.$$

Therefore, by setting $\alpha_k := \frac{1}{n}\mathbf{1}^\top e_k$, we may write

$$e_k = \alpha_k \mathbf{1} + z_k.$$

Then, we have

$$A_{\mu_k} e_k = \gamma \alpha_k \mathbf{1} + A_{\mu_k} z_k.$$

Applying $\mathbf{\Pi}_\perp$ to both sides yields

$$z_{k+1} = \mathbf{\Pi}_\perp e_{k+1} = \mathbf{\Pi}_\perp A_{\mu_k} e_k = \mathbf{\Pi}_\perp A_{\mu_k} z_k = \mathbf{\Pi}_\perp A_{\mu_k} \mathbf{\Pi}_\perp z_k = \bar{A}_{\mu_k} z_k,$$

which proves the first claim. For the second claim, note that $\mathcal{X}_1 = Q^* + \mathrm{span}(\mathbf{1})$. Therefore, we get

$$\mathrm{dist}_2(Q_k, \mathcal{X}_1) = \mathrm{dist}_2(e_k, \mathrm{span}(\mathbf{1})) = \|\mathbf{\Pi}_\perp e_k\|_2 = \|z_k\|_2,$$

which completes the proof. □

D.7 Statement and proof of Lemma 7

Lemma 7. *The following statements hold:*

1. *Let (v, λ) be an eigenvector-eigenvalue pair of A_i. Then*

$$\bar{A}_i(\mathbf{\Pi}_\perp v) = \lambda(\mathbf{\Pi}_\perp v).$$

   In particular, if $\mathbf{\Pi}_\perp v \neq 0$, then $(\mathbf{\Pi}_\perp v, \lambda)$ is an eigenvector-eigenvalue pair of $\bar{A}_i$.

2. *$(\mathbf{1}, 0)$ is an eigenvector-eigenvalue pair of $\bar{A}_i$.*

Proof. Recall that

$$\bar{A}_i = \mathbf{\Pi}_\perp A_i \mathbf{\Pi}_\perp, \qquad \mathbf{\Pi}_\perp = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top, \qquad A_i \mathbf{1} = \gamma \mathbf{1}.$$

To prove the first statement, let (v, λ) be an eigenvector-eigenvalue pair of A_i so that $A_i v = \lambda v$. Since $\mathbf{\Pi}_\perp v = v - \frac{\mathbf{1}^\top v}{n}\mathbf{1}$, we have

$$A_i(\mathbf{\Pi}_\perp v) = A_i v - \frac{\mathbf{1}^\top v}{n} A_i \mathbf{1} = \lambda v - \frac{\gamma\, \mathbf{1}^\top v}{n}\mathbf{1}.$$

Applying $\mathbf{\Pi}_\perp$ to both sides yields

$$\begin{aligned}\bar{A}_i(\mathbf{\Pi}_\perp v) &= \mathbf{\Pi}_\perp A_i \mathbf{\Pi}_\perp v \\ &= \mathbf{\Pi}_\perp \left(\lambda v - \frac{\gamma \mathbf{1}^\top v}{n}\mathbf{1}\right) \\ &= \lambda \mathbf{\Pi}_\perp v - \frac{\gamma \mathbf{1}^\top v}{n}\mathbf{\Pi}_\perp \mathbf{1}.\end{aligned}$$

Since $\mathbf{\Pi}_\perp \mathbf{1} = 0$, it follows that $\bar{A}_i(\mathbf{\Pi}_\perp v) = \lambda \mathbf{\Pi}_\perp v$. Therefore, if $\mathbf{\Pi}_\perp v \neq 0$, then $(\mathbf{\Pi}_\perp v, \lambda)$ is an eigenvector-eigenvalue pair of $\bar{A}_i$. For the second statement, because $\mathbf{\Pi}_\perp \mathbf{1} = 0$, we have $\bar{A}_i \mathbf{1} = \mathbf{\Pi}_\perp A_i \mathbf{\Pi}_\perp \mathbf{1} = \mathbf{\Pi}_\perp A_i 0 = 0$. Therefore, $(\mathbf{1}, 0)$ is an eigenvector-eigenvalue pair of $\bar{A}_i$. This completes the proof. □

D.8 Statement and proof of Lemma 8

Lemma 8. *For any vector $x \in \mathbb{R}^n$, any positive integer k, and any switching sequence $\bar{\sigma}_k := (\sigma_1, \sigma_2, \ldots, \sigma_k) \in \{1, 2, \ldots, M\}^k$, we have*

$$\mathbf{\Pi}_\perp A_{\sigma_k} \cdots A_{\sigma_2} A_{\sigma_1} x = \bar{A}_{\sigma_k} \cdots \bar{A}_{\sigma_2} \bar{A}_{\sigma_1} x.$$

Proof. We prove the claim by induction on $k \geq 0$. First, consider the case $k = 1$. Since $x = \mathbf{\Pi}_\perp x + (I - \mathbf{\Pi}_\perp)x$, we obtain

$$\mathbf{\Pi}_\perp A_{\sigma_1} x = \mathbf{\Pi}_\perp A_{\sigma_1} \mathbf{\Pi}_\perp x + \mathbf{\Pi}_\perp A_{\sigma_1} (I - \mathbf{\Pi}_\perp)x.$$

Now note that $(I - \mathbf{\Pi}_\perp)x \in \operatorname{span}(\mathbf{1})$, and hence, there exists a scalar $c \in \mathbb{R}$ such that

$$(I - \mathbf{\Pi}_\perp)x = c\mathbf{1}.$$

Since $A_i \mathbf{1} = \gamma \mathbf{1}$ for every i, it follows that

$$\mathbf{\Pi}_\perp A_{\sigma_1} (I - \mathbf{\Pi}_\perp)x = c\, \mathbf{\Pi}_\perp A_{\sigma_1} \mathbf{1} = c\gamma \mathbf{\Pi}_\perp \mathbf{1} = 0.$$

Therefore,

$$\mathbf{\Pi}_\perp A_{\sigma_1} x = \mathbf{\Pi}_\perp A_{\sigma_1} \mathbf{\Pi}_\perp x = \bar{A}_{\sigma_1} x.$$

Hence, the claim holds for $k = 1$. Assume now that the identity holds for some $k \geq 1$, namely,

$$\mathbf{\Pi}_\perp A_{\sigma_k} \cdots A_{\sigma_2} A_{\sigma_1} x = \bar{A}_{\sigma_k} \cdots \bar{A}_{\sigma_2} \bar{A}_{\sigma_1} x.$$

We show that it also holds for $k + 1$. We write

$$A_{\sigma_k} \cdots A_{\sigma_1} x = \mathbf{\Pi}_\perp A_{\sigma_k} \cdots A_{\sigma_1} x + (I - \mathbf{\Pi}_\perp) A_{\sigma_k} \cdots A_{\sigma_1} x.$$

Applying $\mathbf{\Pi}_\perp A_{\sigma_{k+1}}$ to both sides gives

$$\begin{aligned}\mathbf{\Pi}_\perp A_{\sigma_{k+1}} A_{\sigma_k} \cdots A_{\sigma_1} x = {} & \mathbf{\Pi}_\perp A_{\sigma_{k+1}} \mathbf{\Pi}_\perp A_{\sigma_k} \cdots A_{\sigma_1} x \\ & + \mathbf{\Pi}_\perp A_{\sigma_{k+1}} (I - \mathbf{\Pi}_\perp) A_{\sigma_k} \cdots A_{\sigma_1} x.\end{aligned}$$

Again, since $(I - \mathbf{\Pi}_\perp) A_{\sigma_k} \cdots A_{\sigma_1} x \in \operatorname{span}(\mathbf{1})$, there exists some scalar $d \in \mathbb{R}$ such that

$$(I - \mathbf{\Pi}_\perp) A_{\sigma_k} \cdots A_{\sigma_1} x = d\mathbf{1}.$$

Using $A_i \mathbf{1} = \gamma \mathbf{1}$ and $\mathbf{\Pi}_\perp \mathbf{1} = 0$, we obtain

$$\mathbf{\Pi}_\perp A_{\sigma_{k+1}} (I - \mathbf{\Pi}_\perp) A_{\sigma_k} \cdots A_{\sigma_1} x = d\, \mathbf{\Pi}_\perp A_{\sigma_{k+1}} \mathbf{1} = d\gamma \mathbf{\Pi}_\perp \mathbf{1} = 0.$$

Hence,

$$\mathbf{\Pi}_\perp A_{\sigma_{k+1}} A_{\sigma_k} \cdots A_{\sigma_1} x = \mathbf{\Pi}_\perp A_{\sigma_{k+1}} \mathbf{\Pi}_\perp A_{\sigma_k} \cdots A_{\sigma_1} x = \bar{A}_{\sigma_{k+1}} \mathbf{\Pi}_\perp A_{\sigma_k} \cdots A_{\sigma_1} x.$$

By the induction hypothesis, $\mathbf{\Pi}_\perp A_{\sigma_k} \cdots A_{\sigma_1} x = \bar{A}_{\sigma_k} \cdots \bar{A}_{\sigma_1} x$. Substituting this into the previous equality yields

$$\mathbf{\Pi}_\perp A_{\sigma_{k+1}} A_{\sigma_k} \cdots A_{\sigma_1} x = \bar{A}_{\sigma_{k+1}} \bar{A}_{\sigma_k} \cdots \bar{A}_{\sigma_1} x.$$

Therefore, the claim holds for $k + 1$. By induction, the result follows for all $k \geq 1$. □

D.9 Statement and proof of Lemma 9

Lemma 9. *Let the eigenvalues of $B_i := \gamma^{-1} A_i = P\Pi^{\pi_i}$, counted with algebraic multiplicity over the complexification, be ordered as*

$$\lambda_{1,i} = 1, \lambda_{2,i}, \ldots, \lambda_{n,i}, \qquad n := |\mathcal{S}||\mathcal{A}|,$$

where $\lambda_{1,i} = 1$ is the Perron eigenvalue associated with the invariant constant direction and the remaining eigenvalues are ordered so that

$$|\lambda_{2,i}| \geq |\lambda_{3,i}| \geq \cdots \geq |\lambda_{n,i}|.$$

If the eigenvalue 1 has algebraic multiplicity larger than one, another copy of 1 may appear among $\lambda_{2,i}, \ldots, \lambda_{n,i}$. Then,

$$\rho(\bar{A}_i) = \gamma |\lambda_{2,i}|,$$

where ρ denotes the spectral radius.

Proof. All spectral statements below are understood over $\mathbb{C}$. Recall that $A_i = \gamma P \Pi^{\pi_i}$ satisfies $A_i \mathbf{1} = \gamma \mathbf{1}$. Let $W := \mathrm{span}(\mathbf{1})^{\perp}$, and choose real vectors $w_2, \ldots, w_n$ forming a basis of W. Define the invertible matrix

$$T := [\mathbf{1} \quad w_2 \quad \cdots \quad w_n] \in \mathbb{R}^{n \times n}.$$

Since $A_i \mathbf{1} = \gamma \mathbf{1}$, the one-dimensional subspace $\mathrm{span}(\mathbf{1})$ is A_i-invariant. Therefore, in the basis induced by T, the matrix A_i has the block upper-triangular form

$$T^{-1} A_i T = \begin{bmatrix} \gamma & \alpha_i^{\top} \\ 0 & \Gamma_i \end{bmatrix}$$

for some vector $\alpha_i \in \mathbb{R}^{n-1}$ and some matrix $\Gamma_i \in \mathbb{R}^{(n-1)\times(n-1)}$. Complexifying this block representation does not change the eigenvalues. Next, since $\mathbf{\Pi}_{\perp}$ is the orthogonal projection onto $W = \mathrm{span}(\mathbf{1})^{\perp}$, we have

$$\mathbf{\Pi}_{\perp} \mathbf{1} = 0, \qquad \mathbf{\Pi}_{\perp} w_j = w_j, \quad j = 2, \ldots, n.$$

Hence, in the same basis,

$$T^{-1} \mathbf{\Pi}_{\perp} T = \begin{bmatrix} 0 & 0 \\ 0 & I_{n-1} \end{bmatrix}.$$

Therefore,

$$\begin{aligned} T^{-1} \bar{A}_i T &= T^{-1} \mathbf{\Pi}_{\perp} A_i \mathbf{\Pi}_{\perp} T \\ &= (T^{-1} \mathbf{\Pi}_{\perp} T)(T^{-1} A_i T)(T^{-1} \mathbf{\Pi}_{\perp} T) \\ &= \begin{bmatrix} 0 & 0 \\ 0 & I_{n-1} \end{bmatrix} \begin{bmatrix} \gamma & \alpha_i^{\top} \\ 0 & \Gamma_i \end{bmatrix} \begin{bmatrix} 0 & 0 \\ 0 & I_{n-1} \end{bmatrix} \\ &= \begin{bmatrix} 0 & 0 \\ 0 & \Gamma_i \end{bmatrix}. \end{aligned}$$

It follows that

$$\sigma(\bar{A}_i) = \{0\} \cup \sigma(\Gamma_i).$$

On the other hand, since

$$T^{-1} A_i T = \begin{bmatrix} \gamma & \alpha_i^{\top} \\ 0 & \Gamma_i \end{bmatrix}$$

is block upper triangular, its spectrum is

$$\sigma(A_i) = \{\gamma\} \cup \sigma(\Gamma_i).$$

Equivalently, if the eigenvalues of $B_i = \gamma^{-1} A_i$ are ordered as in the statement, then the eigenvalues of A_i are $\gamma, \gamma\lambda_{2,i}, \ldots, \gamma\lambda_{n,i}$, and the eigenvalues of $\bar{A}_i$ are $0, \gamma\lambda_{2,i}, \ldots, \gamma\lambda_{n,i}$. Therefore,

$$\rho(\bar{A}_i) = \max\{0, \gamma|\lambda_{2,i}|, \ldots, \gamma|\lambda_{n,i}|\} = \gamma|\lambda_{2,i}|.$$

This completes the proof. ∎

D.10 Statement and proof of Lemma 10

Lemma 10 (A computable upper bound on the JSR)**.** *Let $\bar{\mathcal{H}} := \{\bar{A}_1, \bar{A}_2, \dots, \bar{A}_M\}$ and suppose that there exist a submultiplicative matrix norm $\|\cdot\|_*$ and a constant $\bar{\beta} \in (0,1)$ such that*

$$\|\bar{A}_i\|_* \le \bar{\beta}, \qquad \forall i \in \{1, 2, \dots, M\}.$$

Then the JSR of $\bar{\mathcal{H}}$ satisfies $\rho(\bar{A}_1, \bar{A}_2, \dots, \bar{A}_M) \le \bar{\beta}$.

Proof. By the definition of the JSR, for any positive integer k,

$$\rho(\bar{A}_1, \dots, \bar{A}_M) = \lim_{k\to\infty} \max_{\bar{\sigma}_k \in \{1,\dots,M\}^k} \left\|\bar{A}_{\sigma_k} \cdots \bar{A}_{\sigma_1}\right\|_*^{1/k}.$$

Fix any switching sequence $\bar{\sigma}_k = (\sigma_1, \sigma_2, \dots, \sigma_k) \in \{1, \dots, M\}^k$. Since $\|\cdot\|_*$ is submultiplicative, we have

$$\left\|\bar{A}_{\sigma_k} \cdots \bar{A}_{\sigma_1}\right\|_* \le \prod_{j=1}^{k} \|\bar{A}_{\sigma_j}\|_* \le \bar{\beta}^k.$$

Taking the maximum over all switching sequences yields $\max_{\bar{\sigma}_k \in \{1,\dots,M\}^k} \left\|\bar{A}_{\sigma_k} \cdots \bar{A}_{\sigma_1}\right\|_* \le \bar{\beta}^k$. Taking the k-th root and letting $k \to \infty$, we obtain $\rho(\bar{A}_1, \bar{A}_2, \dots, \bar{A}_M) \le \bar{\beta}$. This completes the proof. □

D.11 Proof of Lemma 5

Proof. Fix any submultiplicative matrix norm $\|\cdot\|$, any positive integer $k \ge 0$, and any switching sequence $\bar{\sigma}_k = (\sigma_1, \dots, \sigma_k) \in \{1, 2, \dots, M\}^k$. Lemma 8 yields

$$\bar{A}_{\sigma_k} \cdots \bar{A}_{\sigma_1} x = \mathbf{\Pi}_\perp A_{\sigma_k} \cdots A_{\sigma_1} x \qquad \forall x \in \mathbb{R}^n.$$

Hence, this implies $\bar{A}_{\sigma_k} \cdots \bar{A}_{\sigma_1} = \mathbf{\Pi}_\perp A_{\sigma_k} \cdots A_{\sigma_1}$. Therefore, it follows that

$$\left\|\bar{A}_{\sigma_k} \cdots \bar{A}_{\sigma_1}\right\| \le \|\mathbf{\Pi}_\perp\| \left\|A_{\sigma_k} \cdots A_{\sigma_1}\right\|.$$

Taking the maximum over all switching sequences gives

$$\max_{\bar{\sigma}_k} \left\|\bar{A}_{\sigma_k} \cdots \bar{A}_{\sigma_1}\right\| \le \|\mathbf{\Pi}_\perp\| \max_{\bar{\sigma}_k} \left\|A_{\sigma_k} \cdots A_{\sigma_1}\right\|.$$

Taking the k-th root yields

$$\left(\max_{\bar{\sigma}_k} \left\|\bar{A}_{\sigma_k} \cdots \bar{A}_{\sigma_1}\right\|\right)^{1/k} \le \|\mathbf{\Pi}_\perp\|^{1/k} \left(\max_{\bar{\sigma}_k} \left\|A_{\sigma_k} \cdots A_{\sigma_1}\right\|\right)^{1/k}.$$

Letting $k \to \infty$, and using $\|\mathbf{\Pi}_\perp\|^{1/k} \to 1$, we obtain

$$\rho(\bar{A}_1, \dots, \bar{A}_M) \le \rho(A_1, \dots, A_M).$$

Since $\rho(A_1, \dots, A_M) = \gamma$ from Lemma 3, it follows that $\rho(\bar{A}_1, \dots, \bar{A}_M) \le \gamma$. This completes the proof. □

D.12 Statement and proof of Lemma 12

Lemma 12 (Common Lyapunov function for the restricted switching family)**.** *Let*

$$\bar{\mathcal{H}} := \{\bar{A}_1, \bar{A}_2, \dots, \bar{A}_M\}, \qquad \bar{\rho} := \rho(\bar{A}_1, \bar{A}_2, \dots, \bar{A}_M),$$

and fix any $\epsilon > 0$ such that

$$\beta_\epsilon := \bar{\rho} + \epsilon \in (0, 1).$$

For each integer $t \ge 0$, define the function

$$V_\varepsilon^t(x) := \sum_{k=0}^{t} \beta_\varepsilon^{-2k} \max_{\bar{\sigma}_k \in \{1,2,\dots,M\}^k} \left\|\bar{A}_{\sigma_k} \cdots \bar{A}_{\sigma_1} x\right\|_2^2, \qquad x \in \mathbb{R}^n.$$

Then the following statements hold:

1. *For every $t \geq 0$,*
$$V_\varepsilon^{t+1}(x) \geq \|x\|_2^2 + \beta_\varepsilon^{-2} \max_{i\in\{1,\dots,M\}} V_\varepsilon^t(\bar{A}_i x).$$

2. *For every $t \geq 0$ and every $x \in \mathbb{R}^n$, $V_\varepsilon^t(\lambda x) = |\lambda|^2 V_\varepsilon^t(x)$ for all $\lambda \in \mathbb{R}$, and $V_\varepsilon^t(x) \leq V_\varepsilon^{t+1}(x)$.*

3. *There exists a constant $C_\varepsilon > 0$ such that*
$$\|x\|_2^2 \leq V_\varepsilon^t(x) \leq C_\varepsilon \|x\|_2^2, \qquad \forall x \in \mathbb{R}^n,\ \forall t \geq 0.$$

4. *For every $x \in \mathbb{R}^n$, the limit*
$$V_\varepsilon^\infty(x) := \lim_{t\to\infty} V_\varepsilon^t(x)$$
exists and is finite. Moreover, $\|x\|_2^2 \leq V_\varepsilon^\infty(x) \leq C_\varepsilon \|x\|_2^2$.

5. *The function $p_\varepsilon(x) := \sqrt{V_\varepsilon^\infty(x)}$ is a norm on $\mathbb{R}^n$.*

6. *The function V_ε^∞ satisfies the Lyapunov inequality*
$$V_\varepsilon^\infty(\bar{A}_i x) \leq \beta_\varepsilon^2 V_\varepsilon^\infty(x), \qquad \forall x \in \mathbb{R}^n,\ \forall i \in \{1,\dots,M\}.$$
Equivalently,
$$p_\varepsilon(\bar{A}_i x) \leq \beta_\varepsilon p_\varepsilon(x), \qquad \forall x \in \mathbb{R}^n,\ \forall i \in \{1,\dots,M\}.$$

7. *Consequently,*
$$\|\bar{A}_i\|_{p_\varepsilon} \leq \beta_\varepsilon, \qquad \forall i \in \{1,\dots,M\},$$
where $\|\bar{A}_i\|_{p_\varepsilon}$ is the induced matrix norm generated by p_ε
$$\|\bar{A}_i\|_{p_\varepsilon} := \sup_{x\neq 0} \frac{p_\varepsilon(\bar{A}_i x)}{p_\varepsilon(x)}.$$
Therefore, we have $\rho(\bar{A}_1, \bar{A}_2, \dots, \bar{A}_M) \leq \beta_\varepsilon$.

Proof. We prove the statements one by one.

Proof of 1). By definition, one has
$$V_\varepsilon^{t+1}(x) = \sum_{k=0}^{t+1} \beta_\varepsilon^{-2k} \max_{\bar{\sigma}_k\in\{1,\dots,M\}^k} \left\|\bar{A}_{\sigma_k}\cdots\bar{A}_{\sigma_1}x\right\|_2^2.$$
The $k=0$ term is simply $\|x\|_2^2$. For $k \geq 1$, writing $k = j+1$, we obtain
$$\begin{aligned}
V_\varepsilon^{t+1}(x) &= \|x\|_2^2 + \sum_{j=0}^{t} \beta_\varepsilon^{-2(j+1)} \max_{\bar{\sigma}_{j+1}} \left\|\bar{A}_{\sigma_{j+1}}\cdots\bar{A}_{\sigma_1}x\right\|_2^2 \\
&= \|x\|_2^2 + \beta_\varepsilon^{-2} \sum_{j=0}^{t} \beta_\varepsilon^{-2j} \max_{i\in\{1,\dots,M\}} \max_{\bar{\tau}_j\in\{1,\dots,M\}^j} \left\|\bar{A}_{\tau_j}\cdots\bar{A}_{\tau_1}\bar{A}_i x\right\|_2^2 \\
&\geq \|x\|_2^2 + \beta_\varepsilon^{-2} \max_{i\in\{1,\dots,M\}} \sum_{j=0}^{t} \beta_\varepsilon^{-2j} \max_{\bar{\tau}_j\in\{1,\dots,M\}^j} \left\|\bar{A}_{\tau_j}\cdots\bar{A}_{\tau_1}\bar{A}_i x\right\|_2^2 \\
&= \|x\|_2^2 + \beta_\varepsilon^{-2} \max_{i\in\{1,\dots,M\}} V_\varepsilon^t(\bar{A}_i x).
\end{aligned}$$
This proves 1).

Proof of 2). The homogeneity follows immediately from the homogeneity of the Euclidean norm:
$$V_\varepsilon^t(\lambda x) = \sum_{k=0}^{t} \beta_\varepsilon^{-2k} \max_{\bar{\sigma}_k} \|\bar{A}_{\sigma_k}\cdots\bar{A}_{\sigma_1}(\lambda x)\|_2^2 = |\lambda|^2 V_\varepsilon^t(x).$$

The monotonicity $V_\varepsilon^t(x) \le V_\varepsilon^{t+1}(x)$ 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: $V_\varepsilon^t(x) \ge \|x\|_2^2$. For the upper bound, since $\bar\rho < \beta_\varepsilon$, choose any number η such that $\bar\rho < \eta < \beta_\varepsilon$. By the definition of the JSR, there exists an integer $K \ge 0$ such that

$$\max_{\bar\sigma_k \in \{1,\dots,M\}^k} \left\|\bar A_{\sigma_k} \cdots \bar A_{\sigma_1}\right\|_2^{1/k} \le \eta, \qquad \forall k \ge K.$$

Hence, for all $k \ge K$, we have

$$\max_{\bar\sigma_k \in \{1,\dots,M\}^k} \left\|\bar A_{\sigma_k} \cdots \bar A_{\sigma_1}\right\|_2 \le \eta^k.$$

Now define

$$C_0 := \max\Big\{1,\ \max_{0\le k\le K-1} \eta^{-k} \max_{\bar\sigma_k} \|\bar A_{\sigma_k} \cdots \bar A_{\sigma_1}\|_2\Big\}.$$

Then, for every $k \ge 0$,

$$\max_{\bar\sigma_k} \left\|\bar A_{\sigma_k} \cdots \bar A_{\sigma_1}\right\|_2 \le C_0\, \eta^k.$$

Therefore, it follows that

$$\begin{aligned} V_\varepsilon^t(x) &= \sum_{k=0}^{t} \beta_\varepsilon^{-2k} \max_{\bar\sigma_k} \left\|\bar A_{\sigma_k} \cdots \bar A_{\sigma_1} x\right\|_2^2 \\ &\le \sum_{k=0}^{t} \beta_\varepsilon^{-2k} \left(\max_{\bar\sigma_k} \left\|\bar A_{\sigma_k} \cdots \bar A_{\sigma_1}\right\|_2\right)^2 \|x\|_2^2 \\ &\le C_0^2 \sum_{k=0}^{t} \left(\frac{\eta}{\beta_\varepsilon}\right)^{2k} \|x\|_2^2 \\ &\le C_0^2 \sum_{k=0}^{\infty} \left(\frac{\eta}{\beta_\varepsilon}\right)^{2k} \|x\|_2^2. \end{aligned}$$

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

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

we obtain $V_\varepsilon^t(x) \le C_\varepsilon \|x\|_2^2$ for all x and $t \ge 0$. This proves 3).

Proof of 4). By 2), the sequence $V_\varepsilon^t(x)$ is nondecreasing in t, and by 3), it is uniformly bounded above by $C_\varepsilon\|x\|_2^2$. Hence the limit $V_\varepsilon^\infty(x) := \lim_{t\to\infty} V_\varepsilon^t(x)$ exists and is finite for every $x \in \mathbb{R}^n$. Passing to the limit in the bounds of 3) yields $\|x\|_2^2 \le V_\varepsilon^\infty(x) \le C_\varepsilon\|x\|_2^2$.

Proof of 5). For each fixed $k \ge 1$, define

$$\nu_k(x) := \beta_\varepsilon^{-k} \max_{\bar\sigma_k \in \{1,\dots,M\}^k} \|\bar A_{\sigma_k} \cdots \bar A_{\sigma_1} x\|_2.$$

Moreover, define $\nu_0(x) := \|x\|_2$. For each $k \ge 0$, ν_k is a seminorm, since it is the pointwise maximum of seminorms. Then

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

Because $\nu_0(x) = \|x\|_2$ is a norm, p_ε^t is a norm for every $t \ge 0$. Indeed, positivity and absolute homogeneity are immediate, and the triangle inequality follows from Minkowski's inequality applied to the vector $(\nu_0(x), \nu_1(x), \dots, \nu_t(x))$. Now, by definition,

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

where the limit is monotone increasing. Since each p_ε^t is a norm, we have for all $x, y \in \mathbb{R}^n$, $p_\varepsilon^t(x+y) \le p_\varepsilon^t(x) + p_\varepsilon^t(y)$. Letting $t \to \infty$, we obtain

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

Absolute homogeneity is inherited from $V_\varepsilon^\infty(\lambda x) = |\lambda|^2 V_\varepsilon^\infty(x)$, and positive definiteness follows from

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

Hence p_ε is a norm.

Proof of 6). Using 1), we have

$$V_\varepsilon^{t+1}(x) \ge \|x\|_2^2 + \beta_\varepsilon^{-2} \max_i V_\varepsilon^t(\bar{A}_i x).$$

Therefore,

$$\max_i V_\varepsilon^t(\bar{A}_i x) \le \beta_\varepsilon^2 \big(V_\varepsilon^{t+1}(x) - \|x\|_2^2\big) \le \beta_\varepsilon^2 V_\varepsilon^{t+1}(x).$$

Fixing i, we have $V_\varepsilon^t(\bar{A}_i x) \le \beta_\varepsilon^2 V_\varepsilon^{t+1}(x)$. Letting $t \to \infty$ and using the monotone convergence of both sides gives

$$V_\varepsilon^\infty(\bar{A}_i x) \le \beta_\varepsilon^2 V_\varepsilon^\infty(x).$$

Taking square roots yields $p_\varepsilon(\bar{A}_i x) \le \beta_\varepsilon p_\varepsilon(x)$.

Proof of 7). From 6), the induced matrix norm generated by p_ε satisfies

$$\|\bar{A}_i\|_{p_\varepsilon} := \sup_{x \ne 0} \frac{p_\varepsilon(\bar{A}_i x)}{p_\varepsilon(x)} \le \beta_\varepsilon, \qquad \forall i \in \{1, \dots, M\}.$$

Hence, for any switching sequence $(\sigma_1, \dots, \sigma_k)$,

$$\|\bar{A}_{\sigma_k} \cdots \bar{A}_{\sigma_1}\|_{p_\varepsilon} \le \prod_{j=1}^{k} \|\bar{A}_{\sigma_j}\|_{p_\varepsilon} \le \beta_\varepsilon^k.$$

Taking the maximum over all switching sequences, then the k-th root, and finally the limit as $k \to \infty$, we obtain $\rho(\bar{A}_1, \bar{A}_2, \dots, \bar{A}_M) \le \beta_\varepsilon$. This completes the proof. □

D.13 Statement and Proof of Lemma 13

Lemma 13 (Convex-hull extension of the Lyapunov function)**.** *Let V_∞^ε be the convex homogeneous Lyapunov function defined in Lemma 12, and fix any $\epsilon > 0$ such that*

$$\beta_\epsilon := \bar{\rho} + \epsilon \in (0, 1).$$

Then V_∞^ε is convex, and for every stochastic policy μ,

$$V_\infty^\varepsilon(\bar{A}_\mu x) \le \beta_\varepsilon^2 V_\infty^\varepsilon(x), \qquad \forall x \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}.$$

Proof. For each deterministic policy $\pi \in \Theta$, let

$$A_\pi := \gamma P \Pi^\pi, \qquad \bar{A}_\pi := \mathbf{\Pi}_\perp A_\pi \mathbf{\Pi}_\perp.$$

Any stochastic policy μ can be represented as a convex combination of deterministic policies:

$$\Pi^\mu = \sum_{\pi \in \Theta} c_\pi(\mu) \Pi^\pi, \qquad c_\pi(\mu) \ge 0, \qquad \sum_{\pi \in \Theta} c_\pi(\mu) = 1,$$

where

$$c_\pi(\mu) := \prod_{s \in \mathcal{S}} \mu(\pi(s)|s).$$

Consequently, we have

$$A_\mu = \sum_{\pi \in \Theta} c_\pi(\mu) A_\pi, \qquad \bar{A}_\mu = \sum_{\pi \in \Theta} c_\pi(\mu) \bar{A}_\pi.$$

Conversely, any convex combination of deterministic-policy matrices is generated by the stochastic policy whose statewise action probabilities are the corresponding marginals,

$$\mu_c(a|s) := \sum_{\pi\in\Theta:\,\pi(s)=a} c_\pi.$$

Thus the stochastic-policy family is exactly the convex hull of the deterministic-policy family; the proof below uses only the preceding inclusion. Now, for each finite t, the function V_t^ε is convex because it is a sum of terms of the form $x \mapsto \max_{\bar\sigma_k} \|\bar A_{\sigma_k}\cdots\bar A_{\sigma_1}x\|_2^2$, which is the pointwise maximum of convex quadratic functions. Since $V_\infty^\varepsilon(x) = \sup_{t\ge 0} V_t^\varepsilon(x)$, it follows that V_∞^ε is also convex. Therefore, using Jensen's inequality leads to

$$\begin{aligned} V_\infty^\varepsilon(\bar A_\mu x) &= V_\infty^\varepsilon\left(\sum_{\pi\in\Theta} c_\pi(\mu)\bar A_\pi x\right) \\ &\le \sum_{\pi\in\Theta} c_\pi(\mu)\, V_\infty^\varepsilon(\bar A_\pi x) \\ &\le \max_{\pi\in\Theta} V_\infty^\varepsilon(\bar A_\pi x). \end{aligned}$$

By Lemma 12, one gets $\max_{\pi\in\Theta} V_\infty^\varepsilon(\bar A_\pi x) \le \beta_\varepsilon^2 V_\infty^\varepsilon(x)$. Combining the two inequalities completes the proof. □

D.14 Proof of Theorem 1

Proof. By Lemma 4, we have $z_{k+1} = \bar A_{\mu_k} z_k$. Hence, by Lemma 13, one can derive

$$V_\infty^\varepsilon(z_{k+1}) \le \beta_\varepsilon^2 V_\infty^\varepsilon(z_k), \qquad \forall k \ge 0.$$

Iterating this inequality gives $V_\infty^\varepsilon(z_k) \le \beta_\varepsilon^{2k} V_\infty^\varepsilon(z_0)$. Now define $p_\varepsilon(x) := \sqrt{V_\infty^\varepsilon(x)}$. By Lemma 12, p_ε is also a norm, and there exists $C_\varepsilon > 0$ such that

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

Thus, with $c_\varepsilon := \sqrt{C_\varepsilon}$, we have

$$\|x\|_2 \le p_\varepsilon(x) \le c_\varepsilon \|x\|_2, \qquad \forall x \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}.$$

Therefore, we have

$$\|z_k\|_2 \le p_\varepsilon(z_k) \le \beta_\varepsilon^k p_\varepsilon(z_0) \le c_\varepsilon \beta_\varepsilon^k \|z_0\|_2.$$

Using Lemma 4 once again,

$$\operatorname{dist}_2(Q_k, \mathcal{X}_1) = \|z_k\|_2, \qquad \operatorname{dist}_2(Q_0, \mathcal{X}_1) = \|z_0\|_2,$$

which proves the claim. □

D.15 Statement and Proof of Corollary 2

Corollary 2 (Fast finite-time identification of the POSS)**.** *Let Assumption 1 hold, fix any $\epsilon > 0$ such that*

$$\beta_\epsilon := \bar\rho + \epsilon \in (0,1),$$

and let $c_\varepsilon > 0$ be the constant from Theorem 1. Define

$$K_{\mathrm{id}} := \begin{cases} 0, & c_\varepsilon \operatorname{dist}_2(Q_0, \mathcal{X}_1) < \dfrac{\bar\Delta}{2}, \\ \left\lfloor \dfrac{\log\left(\dfrac{2c_\varepsilon \operatorname{dist}_2(Q_0,\mathcal{X}_1)}{\bar\Delta}\right)}{-\log\beta_\varepsilon} \right\rfloor + 1, & c_\varepsilon \operatorname{dist}_2(Q_0, \mathcal{X}_1) \ge \dfrac{\bar\Delta}{2}. \end{cases}$$

Then, we have $Q_k \in \mathcal{X}^*$ for all $k \ge K_{\mathrm{id}}$. In particular,

$$\pi_{Q_k}(s) \in \Phi^*(s), \qquad \forall s \in \mathcal{S},\ \forall k \ge K_{\mathrm{id}}.$$

Proof. Write

$$e_k = \alpha_k \mathbf{1} + z_k, \qquad z_k = \mathbf{\Pi}_\perp e_k.$$

Since $\alpha_k \mathbf{1}$ does not change the action ordering, only z_k matters for policy identification. If $\|z_k\|_\infty < \frac{\bar{\Delta}}{2}$, then for every state $s \in \mathcal{S}_{\text{sep}}$ and every action $b \notin \Phi^*(s)$,

$$\begin{aligned}\max_{a \in \Phi^*(s)} Q_k(s,a) - Q_k(s,b) &= V^*(s) - Q^*(s,b) + \max_{a \in \Phi^*(s)} e_k(s,a) - e_k(s,b) \\ &\geq \bar{\Delta}_s + \max_{a \in \Phi^*(s)} z_k(s,a) - z_k(s,b) \\ &\geq \bar{\Delta}_s - 2\|z_k\|_\infty.\end{aligned}$$

Hence, it follows that

$$\max_{a \in \Phi^*(s)} Q_k(s,a) - Q_k(s,b) \geq \bar{\Delta}_s - 2\|z_k\|_\infty \geq \bar{\Delta} - 2\|z_k\|_\infty > 0.$$

Consequently, no non-optimal action can be tie-broken greedy, and therefore, we have $\pi_{Q_k}(s) \in \Phi^*(s)$ for all $s \in \mathcal{S}$. By Equation (4), one gets

$$\|z_k\|_2 = \operatorname{dist}_2(Q_k, \mathcal{X}_1) \leq c_\varepsilon \beta_\varepsilon^k \operatorname{dist}_2(Q_0, \mathcal{X}_1).$$

The definition of K_{id} guarantees

$$c_\varepsilon \beta_\varepsilon^k \operatorname{dist}_2(Q_0, \mathcal{X}_1) < \frac{\bar{\Delta}}{2}, \qquad \forall k \geq K_{\text{id}}.$$

Since $\|z_k\|_\infty \leq \|z_k\|_2$, it follows that $Q_k \in \mathcal{X}^*$ for all $k \geq K_{\text{id}}$. □

D.16 Statement and proof of Theorem 2

Theorem 2 (Two-stage convergence)**.** *Let Assumption 1 hold, and let K_{id} be defined as in Corollary 2. Define*

$$\bar{\mathcal{H}}_* := \{\bar{A}_\pi : \pi \in \Theta^*\}, \qquad \bar{\rho}_* := \rho(\bar{\mathcal{H}}_*),$$

where $\bar{A}_\pi := \mathbf{\Pi}_\perp(\gamma P \Pi^\pi)\mathbf{\Pi}_\perp$ for any $\pi \in \Theta^*$. Then, $\bar{\rho}_* \leq \bar{\rho} \leq \gamma$ and the following statements hold.

1. *For any $\varepsilon > 0$ such that $\beta_\varepsilon := \bar{\rho} + \varepsilon < 1$, there exists a constant $c_\varepsilon > 0$ such that*

   $$\operatorname{dist}_2(Q_k, \mathcal{X}_1) \leq c_\varepsilon \beta_\varepsilon^k \operatorname{dist}_2(Q_0, \mathcal{X}_1), \qquad \forall k \geq 0.$$

   In particular, the convergence toward the affine set $\mathcal{X}_1$, and hence toward the POSS $\mathcal{X}^*$, admits exponential upper bounds at any rate larger than the JSR $\bar{\rho}$ of the full restricted switching family, with the stochastic-policy convex hull handled by the Lyapunov inequality.

2. *For all $k \geq K_{\text{id}}$, there exists a policy $\pi_k \in \Theta^*$ such that $Q_{k+1} - Q^* = A_{\pi_k}(Q_k - Q^*)$.*

3. *For any $\varepsilon_* > 0$ such that $\beta_* := \bar{\rho}_* + \varepsilon_* < 1$, there exists a constant $\tilde{C}_{\varepsilon_*} > 0$ such that*

   $$\|z_{K_{\text{id}}+\ell}\|_2 \leq \tilde{C}_{\varepsilon_*} \beta_*^\ell \|z_{K_{\text{id}}}\|_2, \qquad \forall \ell \geq 0.$$

   Thus, after finite-time identification of $\mathcal{X}^*$, the transverse component admits exponential upper bounds at any rate larger than the JSR $\bar{\rho}_*$ of the restricted optimal family.

4. *If, in addition, $\bar{\rho}_* < \gamma$, then one may choose $\varepsilon_* > 0$ sufficiently small so that $\beta_* = \bar{\rho}_* + \varepsilon_* < \gamma$. In this case, there exists a constant $D_{\varepsilon_*} > 0$ such that*

   $$\|Q_{K_{\text{id}}+\ell} - Q^*\|_2 \leq D_{\varepsilon_*} \gamma^\ell \|Q_{K_{\text{id}}} - Q^*\|_2, \qquad \forall \ell \geq 0.$$

Proof. Write

$$e_k := Q_k - Q^*, \qquad \alpha_k := \frac{1}{n} \mathbf{1}^\top e_k, \qquad z_k := \mathbf{\Pi}_\perp e_k, \qquad n := |\mathcal{S}||\mathcal{A}|.$$

The global convergence estimate follows directly from Theorem 1 applied to the full restricted switching family $\bar{A} = \{\bar{A}_1, \ldots, \bar{A}_M\}$. Since $\mathcal{X}_1 \subset \mathcal{X}^*$ and Corollary 2 gives finite-time entrance

into $\mathcal{X}^*$, this proves the first claim and shows that the convergence toward the POSS is controlled by any exponential rate larger than $\bar{\rho}$. Next, by Corollary 2, for all $k \geq K_{\mathrm{id}}$, $Q_k \in \mathcal{X}^*$. Hence, for each such k, the tie-broken greedy policy $\pi_k := \pi_{Q_k}$ belongs to Θ^*. Therefore, $Q_{k+1} = R + \gamma P \Pi^{\pi_k} Q_k$. Since π_k is optimal, it also satisfies $R + \gamma P \Pi^{\pi_k} Q^* = Q^*$. Subtracting the two equations yields $Q_{k+1} - Q^* = A_{\pi_k}(Q_k - Q^*)$, where $A_{\pi_k} := \gamma P \Pi^{\pi_k}$. This proves the second claim. Now define $e_k := Q_k - Q^*$ and $z_k := \mathbf{\Pi}_\perp e_k$. Then, for all $k \geq K_{\mathrm{id}}$,

$$z_{k+1} = \mathbf{\Pi}_\perp e_{k+1} = \mathbf{\Pi}_\perp A_{\pi_k} e_k.$$

Since

$$e_k = \mathbf{\Pi}_\perp e_k + (I - \mathbf{\Pi}_\perp) e_k$$

and $(I - \mathbf{\Pi}_\perp)e_k \in \mathrm{span}(\mathbf{1})$, there exists a scalar $c_k \in \mathbb{R}$ such that

$$(I - \mathbf{\Pi}_\perp) e_k = c_k \mathbf{1}.$$

Using $A_{\pi_k}\mathbf{1} = \gamma \mathbf{1}$ and $\mathbf{\Pi}_\perp \mathbf{1} = 0$, we obtain

$$\mathbf{\Pi}_\perp A_{\pi_k}(I - \mathbf{\Pi}_\perp) e_k = c_k \mathbf{\Pi}_\perp A_{\pi_k} \mathbf{1} = c_k \gamma \mathbf{\Pi}_\perp \mathbf{1} = 0.$$

Therefore,

$$z_{k+1} = \mathbf{\Pi}_\perp A_{\pi_k} \mathbf{\Pi}_\perp e_k = \bar{A}_{\pi_k} z_k,$$

where $\bar{A}_{\pi_k} := \mathbf{\Pi}_\perp A_{\pi_k} \mathbf{\Pi}_\perp$. Thus, after time K_{id}, the projected error evolves according to the switching family $\bar{\mathcal{H}}_* = \{\bar{A}_\pi : \pi \in \Theta^*\}$. Since $\bar{\mathcal{H}}_* \subset \bar{\mathcal{H}}$, monotonicity of the JSR yields $\bar{\rho}_* \leq \bar{\rho}$. On the other hand, Lemma 5 gives $\bar{\rho} \leq \gamma$. Hence, $\bar{\rho}_* \leq \bar{\rho} \leq \gamma$. Fix any $\varepsilon_* > 0$ such that $\beta_* := \bar{\rho}_* + \varepsilon_* < 1$. Applying Lemma 12 to the restricted optimal family $\bar{\mathcal{H}}_*$, there exists a common Lyapunov function for this family. Hence there exists a constant $\tilde{C}_{\varepsilon_*} > 0$ such that

$$\|z_{K_{\mathrm{id}}+\ell}\|_2 \leq \tilde{C}_{\varepsilon_*} \beta_*^\ell \|z_{K_{\mathrm{id}}}\|_2, \qquad \forall \ell \geq 0.$$

This proves the third claim.

It remains to control the component along $\mathbf{1}$. Write

$$e_k = \alpha_k \mathbf{1} + z_k, \qquad \alpha_k := \frac{1}{n} \mathbf{1}^\top e_k.$$

Then

$$\alpha_{k+1} = \frac{1}{n} \mathbf{1}^\top e_{k+1} = \frac{1}{n} \mathbf{1}^\top A_{\pi_k} e_k = \frac{1}{n} \mathbf{1}^\top A_{\pi_k} (\alpha_k \mathbf{1} + z_k).$$

Since $A_{\pi_k}\mathbf{1} = \gamma \mathbf{1}$, we get

$$\alpha_{k+1} = \gamma \alpha_k + \eta_k,$$

where $\eta_k := \frac{1}{n} \mathbf{1}^\top A_{\pi_k} z_k$. Define $L_* := \max_{\pi \in \Theta^*} \frac{1}{n} \|\mathbf{1}^\top A_\pi\|_2$. Then, $|\eta_k| \leq L_* \|z_k\|_2$ so that

$$|\alpha_{k+1}| \leq \gamma |\alpha_k| + |\eta_k| \leq \gamma |\alpha_k| + L_* \|z_k\|_2.$$

Assume now that $\bar{\rho}_* < \gamma$. Then we may choose $\varepsilon_* > 0$ sufficiently small so that

$$\beta_* = \bar{\rho}_* + \varepsilon_* < \gamma.$$

Iterating the recursion for α_k from K_{id}, we obtain

$$|\alpha_{K_{\mathrm{id}}+\ell}| \leq \gamma^\ell |\alpha_{K_{\mathrm{id}}}| + L_* \sum_{j=0}^{\ell-1} \gamma^{\ell-1-j} \|z_{K_{\mathrm{id}}+j}\|_2.$$

Using the bound on z_k gives

$$|\alpha_{K_{\mathrm{id}}+\ell}| \leq \gamma^\ell |\alpha_{K_{\mathrm{id}}}| + L_* \tilde{C}_{\varepsilon_*} \sum_{j=0}^{\ell-1} \gamma^{\ell-1-j} \beta_*^j \|z_{K_{\mathrm{id}}}\|_2.$$

Since $\beta_* < \gamma$,

$$\sum_{j=0}^{\ell-1} \gamma^{\ell-1-j} \beta_*^j = \frac{\gamma^\ell - \beta_*^\ell}{\gamma - \beta_*} \leq \frac{\gamma^\ell}{\gamma - \beta_*}.$$

Hence,

$$|\alpha_{K_{\mathrm{id}}+\ell}| \le \gamma^\ell |\alpha_{K_{\mathrm{id}}}| + \frac{L_* \widetilde{C}_{\varepsilon_*}}{\gamma - \beta_*} \gamma^\ell \|z_{K_{\mathrm{id}}}\|_2.$$

Finally, since $e_{K_{\mathrm{id}}+\ell} = \alpha_{K_{\mathrm{id}}+\ell}\mathbf{1} + z_{K_{\mathrm{id}}+\ell}$, the triangle inequality yields

$$\begin{aligned}
\|e_{K_{\mathrm{id}}+\ell}\|_2 &\le \|\alpha_{K_{\mathrm{id}}+\ell}\mathbf{1}\|_2 + \|z_{K_{\mathrm{id}}+\ell}\|_2 \\
&= \sqrt{n}\,|\alpha_{K_{\mathrm{id}}+\ell}| + \|z_{K_{\mathrm{id}}+\ell}\|_2 \\
&\le \sqrt{n}\,\gamma^\ell \left(|\alpha_{K_{\mathrm{id}}}| + \frac{L_* \widetilde{C}_{\varepsilon_*}}{\gamma - \beta_*} \|z_{K_{\mathrm{id}}}\|_2 \right) + \widetilde{C}_{\varepsilon_*} \beta_*^\ell \|z_{K_{\mathrm{id}}}\|_2 \\
&\le \gamma^\ell \left[\sqrt{n}\,|\alpha_{K_{\mathrm{id}}}| + \left(\frac{\sqrt{n}\,L_* \widetilde{C}_{\varepsilon_*}}{\gamma - \beta_*} + \widetilde{C}_{\varepsilon_*} \right) \|z_{K_{\mathrm{id}}}\|_2 \right],
\end{aligned}$$

where in the last step we used $\beta_*^\ell \le \gamma^\ell$. Finally, we bound $|\alpha_{K_{\mathrm{id}}}|$ and $\|z_{K_{\mathrm{id}}}\|_2$ by $\|e_{K_{\mathrm{id}}}\|_2$. By the definition of $\alpha_{K_{\mathrm{id}}}$,

$$\sqrt{n}\,|\alpha_{K_{\mathrm{id}}}| = \sqrt{n} \left| \frac{1}{n} \mathbf{1}^\top e_{K_{\mathrm{id}}} \right| = \frac{1}{\sqrt{n}} |\mathbf{1}^\top e_{K_{\mathrm{id}}}| \le \frac{1}{\sqrt{n}} \|\mathbf{1}\|_2 \, \|e_{K_{\mathrm{id}}}\|_2 = \|e_{K_{\mathrm{id}}}\|_2.$$

Moreover, since $z_{K_{\mathrm{id}}} = \mathbf{\Pi}_\perp e_{K_{\mathrm{id}}}$ and $\mathbf{\Pi}_\perp$ is an orthogonal projection, $\|z_{K_{\mathrm{id}}}\|_2 \le \|e_{K_{\mathrm{id}}}\|_2$. Substituting these two estimates into the previous bound, we conclude that

$$\|e_{K_{\mathrm{id}}+\ell}\|_2 \le D_{\varepsilon_*} \gamma^\ell \|e_{K_{\mathrm{id}}}\|_2, \qquad \forall \ell \ge 0,$$

where one may take, for instance,

$$D_{\varepsilon_*} := 1 + \widetilde{C}_{\varepsilon_*} + \frac{\sqrt{n}\,L_* \widetilde{C}_{\varepsilon_*}}{\gamma - \beta_*}.$$

Recalling that $e_k = Q_k - Q^*$, we finally obtain

$$\|Q_{K_{\mathrm{id}}+\ell} - Q^*\|_2 = \|e_{K_{\mathrm{id}}+\ell}\|_2 \le D_{\varepsilon_*} \gamma^\ell \|Q_{K_{\mathrm{id}}} - Q^*\|_2, \qquad \forall \ell \ge 0.$$

This completes the proof. □

D.17 Statement and proof of Corollary 3

Corollary 3 (Unique optimal policy case)**.** *Suppose that the optimal policy is unique, namely, $\Theta^* = \{\pi^*\}$. Under the assumptions of Theorem 2, we then have*

$$\pi_{Q_k} = \pi^*, \qquad \forall k \ge K_{\mathrm{id}}.$$

Consequently, for all $k \ge K_{\mathrm{id}}$, $Q_{k+1} - Q^* = A_{\pi^*}(Q_k - Q^*)$, and $z_{k+1} = \bar{A}_{\pi^*} z_k$, where $\bar{A}_{\pi^*} := \mathbf{\Pi}_\perp(\gamma P \Pi^{\pi^*})\mathbf{\Pi}_\perp$. Moreover,

$$\rho(\bar{A}_{\pi^*}) = \gamma |\lambda_2(P\Pi^{\pi^*})|.$$

Hence, in the unique-optimal-policy case, the post-identification transverse dynamics are governed by a single linear system rather than a switching family.

Proof. Since the optimal policy is unique, for each state $s \in \mathcal{S}$ the set $\Phi^*(s)$ is a singleton, namely, $\Phi^*(s) = \{\pi^*(s)\}$. By Corollary 2, we have $Q_k \in \mathcal{X}^*$ for all $k \ge K_{\mathrm{id}}$. Therefore, $\pi_{Q_k}(s) \in \Phi^*(s) = \{\pi^*(s)\}$ for all $s \in \mathcal{S}$, which implies $\pi_{Q_k} = \pi^*$ for all $k \ge K_{\mathrm{id}}$. The stated dynamics then follow immediately from Theorem 2, and the identity $\rho(\bar{A}_{\pi^*}) = \gamma |\lambda_2(P\Pi_{\pi^*})|$ follows from Lemma 9. □

D.18 Proof of Lemma 14

Proof. The bounds $0 \le \tau_{\mathrm{D}}(B) \le 1$, the equivalence between $\tau_{\mathrm{D}}(B) < 1$ and the scrambling property, the submultiplicativity $\tau_{\mathrm{D}}(BC) \le \tau_{\mathrm{D}}(B)\tau_{\mathrm{D}}(C)$, and the exact diameter-seminorm characterization

$$\tau_{\mathrm{D}}(B) = \sup_{v \notin \mathrm{span}(\mathbf{1})} \frac{\|Bv\|_{\mathrm{dm}}}{\|v\|_{\mathrm{dm}}}$$

are standard properties of coefficients of ergodicity for finite stochastic matrices Dobrushin (1956); Hajnal & Bartlett (1958); Seneta (2006). The convexity of τ_{D} follows directly from its definition as a maximum of total-variation distances between rows. Finally, the diameter seminorm is invariant under additive shifts by constants, which gives $\|v + c\mathbf{1}\|_{\mathrm{dm}} = \|v\|_{\mathrm{dm}}$. Since B is row-stochastic, $B\mathbf{1} = \mathbf{1}$. Therefore the projection only removes or adds components in $\mathrm{span}(\mathbf{1})$, and hence

$$\|\mathbf{\Pi}_{\perp} B \mathbf{\Pi}_{\perp} v\|_{\mathrm{dm}} = \|B \mathbf{\Pi}_{\perp} v\|_{\mathrm{dm}}.$$

Taking the supremum over nonconstant v gives the last equality. This completes the proof. □

D.19 Proof of Proposition 3

Proof. Let

$$\bar{\mathcal{B}} := \{\bar{B}_\pi : \pi \in \Theta\}, \qquad \bar{B}_\pi := \mathbf{\Pi}_{\perp} B_\pi \mathbf{\Pi}_{\perp}.$$

Since $\bar{A}_\pi = \gamma \bar{B}_\pi$, we have

$$\bar{\rho} = \gamma \rho(\bar{\mathcal{B}}).$$

The only subtle point is that $\|\cdot\|_{\mathrm{dm}}$ is not a norm on the full space $\mathbb{R}^{|\mathcal{Y}|}$, but it becomes a genuine norm once the constant direction is removed. Let

$$W := \mathrm{span}(\mathbf{1})^{\perp}, \qquad \|w\|_{\mathrm{dm}} := \max_{x \in \mathcal{Y}} w_x - \min_{x \in \mathcal{Y}} w_x, \quad w \in W.$$

On W, $\|\cdot\|_{\mathrm{dm}}$ is a norm. For each $\pi \in \Theta$, define

$$T_\pi := \bar{B}_\pi|_W : W \to W.$$

We identify $\bar{B}_\pi$ with its restriction T_π on W. Since $\bar{B}_\pi \mathbf{1} = 0$ and $\bar{B}_\pi(\mathbb{R}^{|\mathcal{Y}|}) \subset W$, each $\bar{B}_\pi$ has the block form

$$\bar{B}_\pi \sim \begin{bmatrix} 0 & 0 \\ 0 & T_\pi \end{bmatrix} \qquad \text{with respect to } \mathbb{R}^{|\mathcal{Y}|} = \mathrm{span}(\mathbf{1}) \oplus W.$$

Thus, the JSR of the full matrices $\bar{B}_\pi$ equals the JSR of the restricted operators T_π on W.

Now fix $\omega = (\pi_0, \ldots, \pi_{k-1}) \in \Theta^k$. By the same projection-product identity as in Lemma 8,

$$T_{\pi_{k-1}} \cdots T_{\pi_0} = (\mathbf{\Pi}_{\perp} B_{\pi_{k-1}} \cdots B_{\pi_0} \mathbf{\Pi}_{\perp})\big|_W = (\mathbf{\Pi}_{\perp} B_\omega \mathbf{\Pi}_{\perp})\big|_W.$$

By Lemma 14(5)–(6), the induced operator norm of this map on $(W, \|\cdot\|_{\mathrm{dm}})$ is exactly $\tau_{\mathrm{D}}(B_\omega)$. Therefore,

$$\|T_{\pi_{k-1}} \cdots T_{\pi_0}\|_{\mathrm{dm}\to\mathrm{dm}} = \tau_{\mathrm{D}}(B_\omega),$$

and hence

$$\rho(\bar{\mathcal{B}}) = \lim_{k\to\infty} \max_{\omega \in \Theta^k} \tau_{\mathrm{D}}(B_\omega)^{1/k}.$$

We first prove the equivalence between items 1) and 3). Suppose that $\bar{\rho} < \gamma$. Then

$$\rho(\bar{\mathcal{B}}) < 1.$$

Choose β such that $\rho(\bar{\mathcal{B}}) < \beta < 1$. By the definition of the JSR on the normed space $(W, \|\cdot\|_{\mathrm{dm}})$, there exists an integer $L \geq 1$ such that

$$\max_{\omega \in \Theta^L} \|T_\omega\|^{1/L}_{\mathrm{dm}\to\mathrm{dm}} \leq \beta.$$

Using $\|T_\omega\|_{\mathrm{dm}\to\mathrm{dm}} = \tau_{\mathrm{D}}(B_\omega)$, we obtain

$$\max_{\omega \in \Theta^L} \tau_{\mathrm{D}}(B_\omega) \leq \beta^L < 1,$$

which proves item 3).

Conversely, suppose that item 3) holds for some $L \geq 1$, and define

$$\theta_L := \max_{\omega \in \Theta^L} \tau_{\mathrm{D}}(B_\omega) < 1.$$

Then every length-L product of the family $\{T_\pi : \pi \in \Theta\}$ has operator norm at most θ_L. By submultiplicativity, every length-mL product has operator norm at most θ_L^m. Hence

$$\rho(\bar{\mathcal{B}}) \le \theta_L^{1/L} < 1.$$

Multiplying by γ, we obtain $\bar{\rho} < \gamma$. This proves item 1).

The equivalence between items 2) and 3) follows from Lemma 14: for each row-stochastic matrix B, $\tau_{\mathrm{D}}(B) < 1$ if and only if B is scrambling. Since Θ^L is finite, all length-L deterministic products are scrambling if and only if

$$\max_{\omega \in \Theta^L} \tau_{\mathrm{D}}(B_\omega) < 1.$$

We next relate the deterministic and stochastic formulations. For stochastic policies $\mu_0, \dots, \mu_{L-1}$, each B_{μ_j} is a convex combination of deterministic-policy matrices:

$$B_{\mu_j} = \sum_{\pi_j \in \Theta} c_{\pi_j}(\mu_j) B_{\pi_j}.$$

Therefore,

$$B_{\mu_{L-1}} \cdots B_{\mu_0} = \sum_{\omega \in \Theta^L} C_\omega B_\omega,$$

where $C_\omega \ge 0$ and $\sum_{\omega \in \Theta^L} C_\omega = 1$. By the convexity of τ_{D},

$$\tau_{\mathrm{D}}\big(B_{\mu_{L-1}} \cdots B_{\mu_0}\big) \le \sum_{\omega \in \Theta^L} C_\omega \tau_{\mathrm{D}}(B_\omega) \le \max_{\omega \in \Theta^L} \tau_{\mathrm{D}}(B_\omega).$$

Thus item 3) implies item 4). The reverse implication is immediate because deterministic policies are special cases of stochastic policies. Hence items 3) and 4) are equivalent.

Finally, if the equivalent conditions hold, then by item 2) every B_ω, $\omega \in \Theta^L$, is scrambling. Since the family $\{B_\omega : \omega \in \Theta^L\}$ is finite, the quantity

$$\eta_L := \min_{\omega \in \Theta^L} \min_{x,y \in \mathcal{Y}} \sum_{v \in \mathcal{Y}} \min\{(B_\omega)_{xv}, (B_\omega)_{yv}\}$$

is strictly positive. By Definition 3,

$$\tau_{\mathrm{D}}(B_\omega) \le 1 - \eta_L, \qquad \forall \omega \in \Theta^L.$$

Hence every length-L product has operator norm at most $1 - \eta_L$, and

$$\rho(\bar{\mathcal{B}}) \le (1 - \eta_L)^{1/L}.$$

Therefore,

$$\bar{\rho} = \gamma \rho(\bar{\mathcal{B}}) \le \gamma (1 - \eta_L)^{1/L} < \gamma.$$

This completes the proof. □