IS CONSTANT STEP-SIZE DOUBLE Q-LEARNING STABLE? ANALYSIS AND AN ASYMMETRIC VARIANT

Anonymous authors
Paper under double-blind review

ABSTRACT

Double Q-learning is a classical reinforcement learning method designed to reduce maximization bias, yet its worst-case finite-time behavior under constant step sizes has received relatively limited theoretical attention. This paper develops a direct switching-system framework for analyzing the finite-time behavior of constant step-size double Q-learning. By formulating the coupled updates as a switched linear system on an augmented error space, we show that, unlike standard Q-learning, joint spectral radius (JSR) stability of the resulting augmented switching family is not automatically guaranteed. Accordingly, we characterize a mode-wise necessary condition for JSR stability and show that the augmented double-Q switching family can contain unstable modes even for arbitrarily small positive constant step sizes. To address this limitation, we propose a novel asymmetric double Q-learning variant whose tracking structure admits a JSR stability guarantee under explicit constant step-size conditions. We further validate the proposed method through empirical evaluations.

1 INTRODUCTION

Q-learning (Watkins & Dayan, 1992) is a foundational algorithm in reinforcement learning (RL) (Sutton & Barto, 1998), but the maximization step in its Bellman target can introduce positive bias when action-value estimates are noisy, a phenomenon known as maximization bias. Double Q-learning (Hasselt, 2010) addresses this issue by maintaining two Q-tables. Thus, one table selects the greedy action and the other table evaluates it. This decoupling reduces maximization bias. Since its introduction, double Q-learning has been widely adopted in practice, where a constant step-size is frequently employed due to its empirical convenience (Fan et al., 2025; Patil et al., 2025; Du et al., 2026; Zareer & Selmic, 2025). However, despite its widespread practical adoption, theoretical analyses of double Q-learning remain relatively limited compared with those of standard Q-learning (Hasselt, 2010; Xiong et al., 2020; Zhao et al., 2021). This disparity largely stems from the intricate, coupled stochastic recursion introduced to mitigate maximization bias. Furthermore, existing finite-time analyses predominantly rely on expected stochastic behavior, offering no theoretical guarantees for worst-case convergence. For safety-critical autonomous systems, worst-case stability guarantees are of practical importance. Motivated by the lack of worst-case convergence guarantees for double Q-learning, this paper studies double Q-learning from a switching-system viewpoint (Liberzon, 2003). In particular, we focus on the practically important constant step-size regime and develop a framework that enables worst-case stability analysis through switching-system theory. The key observation underlying this formulation is that, in the two-table recursion, each Q-table has its own greedy policy, and the resulting policy-dependent terms act as switching modes. The main technical difficulty in analyzing double Q-learning as a switching-system is that the cross-evaluation target contains a Bellman maximization error, which introduces a nonlinear coupling between the two Q-tables. We remove this nonlinearity exactly using a stochastic-policy linearization: the difference between the maximal value and the optimal value can be written as an average of action-wise errors under a suitable stochastic policy. After this transformation, double Q-learning admits a direct switched linear conditional-mean representation on an augmented error state consisting of the estimation errors of the two Q-tables relative to the optimal Q-function. However, the corresponding switching family is fundamentally different from the switching family of standard Q-learning, because its diagonal blocks include signed policy-difference terms. Consequently, the standard stability argument for Q-learning based on a uniform row-sum bound (Lee, 2026) is no longer directly

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.

applicable. Stability must instead be certified through a joint spectral radius (JSR) condition, an extremal or JSR-induced Lyapunov function, a block common quadratic Lyapunov function, or another suitable switched-system certificate. Without such a certificate, we show that worst-case convergence of double Q-learning is no longer automatically guaranteed under a constant step-size regime. To overcome this limitation, we impose an asymmetric update structure: one table performs the Bellman update, while the other tracks it through a separate tracking recursion. This tracking mechanism yields a common weighted block-norm contraction. Based on this design, the proposed algorithm admits an explicit JSR stability certificate under constant step sizes.

Based on these theoretical and algorithmic developments, our main contributions are summarized as follows. First, we derive an exact direct block switching representation for double Q-learning. Second, we define the augmented double-Q switching family and establish its JSR as a sufficient worst-case stability certificate. Third, we prove finite-time bounds under an explicit JSR stability condition. Fourth, we provide a necessary mode-wise step-size condition and show that augmented double-Q stability is not an automatic consequence of the model. Finally, we propose a novel variant called Asymmetric Tracking double Q-learning (AT-DQ). By introducing an asymmetric tracking update, AT-DQ admits an explicit JSR stability certificate under constant step sizes and is therefore mathematically guaranteed to achieve worst-case stability. We further validate its practical effectiveness through experimental evaluations.

2 Related Works

Despite its widespread empirical adoption, theoretical analysis of double Q-learning remains limited compared to standard Q-learning. The original work of Hasselt (2010) introduced double Q-learning and established its asymptotic convergence. Subsequently, finite-time bounds were derived for diminishing step sizes (Xiong et al., 2020) and constant step sizes (Zhao et al., 2021). However, these analyses focus primarily on expected stochastic behavior and do not directly characterize worst-case switching behavior, which can be relevant in safety-critical applications. Recently, switching-system theory has been employed as a powerful tool for analyzing the convergence and stability properties of Q-learning algorithms (Lee, 2024). Recently, this framework was applied to analyze the finite-time bounds of a double Q-learning variant in Na & Lee (2026). Nevertheless, due to the technical difficulty of handling affine switching terms, their approach does not analyze the true dynamics directly, but rather relies on an auxiliary system that bounds the original dynamics. More recently, Lee (2026) overcame this structural limitation for standard Q-learning, demonstrating how to directly analyze its exact switching dynamics without the aid of auxiliary bounds. Building on these theoretical foundations, this paper directly analyzes the exact switching dynamics of double Q-learning. Rather than directly extending existing analytical frameworks, we show that uniform JSR stability of the augmented standard double-Q switching family is not automatic under constant step sizes, and propose a novel algorithmic variant that admits an explicit JSR stability certificate under constant step sizes.

3 Preliminaries

3.1 Notation

The set of real numbers is denoted by $\mathbb{R}$. For a matrix H, $H^\top$ denotes its transpose. For symmetric matrices, $H \succ 0$ and $H \succeq 0$ denote positive definiteness and positive semidefiniteness, and $H \preceq B$ means $B - H \succeq 0$. The identity matrix is denoted by I. For vectors, e_i is the ith standard basis vector, with dimension clear from context, and $\mathbf{1}$ is the all-ones vector. The quantities $\lambda_{\min}(H)$ and $\lambda_{\max}(H)$ denote the minimum and maximum eigenvalues of a symmetric matrix H. For a square matrix H, $\rho(H)$ is its spectral radius. For a finite set $\mathcal{S}$, $|\mathcal{S}|$ denotes its cardinality. We write

$$\Delta_m := \left\{ q \in \mathbb{R}^m : q_i \geq 0,\ \sum_{i=1}^{m} q_i = 1 \right\}$$

for the probability simplex in $\mathbb{R}^m$. For a finite matrix family $\mathcal{H} = \{H_1, \ldots, H_N\}$,

$$\mathrm{co}(\mathcal{H}) := \left\{ \sum_{i=1}^{N} \lambda_i H_i : \lambda_i \geq 0,\ \sum_{i=1}^{N} \lambda_i = 1 \right\}$$

denotes its convex hull.

3.2 Joint spectral radius and switching-systems

Consider a discrete-time switched linear system

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

where $\mathcal{H}$ is a prescribed matrix family and σ_k is a switching signal. The worst-case exponential growth rate is characterized by the joint spectral radius (JSR) (Rota & Strang, 1960; Jungers, 2009; Tsitsiklis & Blondel, 1997).

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

$$\rho(\mathcal{H}) := \lim_{k\to\infty} \sup_{H_1,\dots,H_k\in\mathcal{H}} \|H_k \cdots H_1\|^{1/k},$$

where the value is independent of the chosen submultiplicative norm. When $\mathcal{H}$ is finite, the supremum for each fixed product length is a maximum over products generated by matrices in $\mathcal{H}$.

For a finite family, passing to the convex hull does not change the JSR (Jungers, 2009):

$$\rho(\mathrm{co}(\mathcal{H})) = \rho(\mathcal{H}).$$

This fact is used repeatedly below to certify deterministic extreme modes while controlling stochastic-policy convex combinations.

3.3 Discounted Markov decision Processes and Bellman operator

Let us consider a finite discounted Markov decision process (MDP) (Puterman, 2014) with state space $\mathcal{S} = \{1,\dots,|\mathcal{S}|\}$, action space $\mathcal{A} = \{1,\dots,|\mathcal{A}|\}$, transition probability $P(s'|s,a)$ for $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, one-step reward $r(s,a,s')$ for $(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}$, expected reward $R(s,a) := \sum_{s'\in\mathcal{S}} P(s'|s,a)r(s,a,s')$, $(s,a) \in \mathcal{S}\times\mathcal{A}$, and discount factor $\gamma \in (0,1)$. Let Θ denote the set of deterministic stationary policies $\pi : \mathcal{S} \to \mathcal{A}$. A Q-function is viewed as a vector $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ using a fixed ordering of state-action pairs. We use the compact notation

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

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

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

Then $P\Pi^\mu \in \mathbb{R}^{(|\mathcal{S}||\mathcal{A}|)\times(|\mathcal{S}||\mathcal{A}|)}$ is the state-action transition matrix under μ. For a deterministic policy $\pi \in \Theta$ we use the same notation Π^π by identifying $\pi(s)$ with its one-hot encoding for each $s \in \mathcal{S}$. For $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, let us define $V_Q(s) := \max_{a\in\mathcal{A}} Q(s,a)$, $s \in \mathcal{S}$ and $V_Q := (V_Q(1),\dots,V_Q(|\mathcal{S}|))^\top \in \mathbb{R}^{|\mathcal{S}|}$. The Bellman optimality operator is $F(Q) := R + \gamma P V_Q$. The optimal Q-function Q^* is the unique fixed point of F, i.e., $Q^* = F(Q^*)$ and $V^* := V_{Q^*}$.

3.4 Stochastic-policy linearization of the Bellman maximum

A central difficulty in the switching-system analysis of Q-learning and double Q-learning is the Bellman maximization operator, which introduces a piecewise-linear nonlinearity into the error dynamics. To resolve this nonlinearity, the following lemma provides an exact stochastic-policy linearization of the Bellman maximum. This identity plays a fundamental role throughout the paper.

Lemma 1. *(Lee, 2026) For every $Q \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$, there exists a stochastic policy $\mu_Q : \mathcal{S} \to \Delta_{|\mathcal{A}|}$ such that*

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

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

The proof is given in Section B.1 of Appendix.

3.5 ASSUMPTION AND DEFINITIONS

Throughout, we adopt the following standard sampling assumptions. Under a fixed behavior policy b, the state–action pair (s, a) is sampled independently at each iteration from the stationary distribution induced by b, yielding the joint distribution $d(s, a) = p(s)b(a \mid s),\ (s, a) \in \mathcal{S} \times \mathcal{A}$. We also impose the following assumptions.

Assumption 1. *(i) $d(s, a) > 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}$; and (ii) the step-size is a constant $\alpha \in (0, 1)$.*

For notational convenience, we define the following quantities that will be used throughout the paper.

Definition 2. *1) Uniform reward bound:* $R_{\max} := \max_{(s,a,s') \in \mathcal{S}\times\mathcal{A}\times\mathcal{S}} |r(s, a, s')|$.

2) Maximum state-action occupancy frequency: $d_{\max} := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} d(s, a)$.

3) Minimum state-action occupancy frequency: $d_{\min} := \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} d(s, a)$.

Since $\mathcal{S}$ and $\mathcal{A}$ are finite, $0 < d_{\min} \le d_{\max} < 1$. Moreover, it is well-known that the optimal Q-function then satisfies $\|Q^*\|_\infty \le R_{\max}/(1 - \gamma)$.

4 DOUBLE Q-LEARNING

We consider the random-table version of double Q-learning, which maintains two Q-tables $Q^1_k, Q^2_k \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Independently, a table selector $J_k \in \{1, 2\}$ is drawn with

$$\mathbb{P}(J_k = 1) = p_1, \qquad \mathbb{P}(J_k = 2) = p_2, \qquad p_1, p_2 > 0, \qquad p_1 + p_2 = 1.$$

To express the double Q-learning recursion compactly, we first introduce some notation. Let $D := \operatorname{diag}\big(d(s, a)\big)_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, where $d(s, a)$ denotes the sampling probability of the state–action pair (s, a). In addition, for each table $X \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, let π_X denote a deterministic greedy policy satisfying $\pi_X(s) \in \arg\max_{a\in\mathcal{A}} X(s, a),\ s \in \mathcal{S}$, where ties are resolved according to a fixed measurable tie-breaking rule. Using this notation, the double Q-learning updates are defined as follows. If $J_k = 1$, only Q^1 is updated:

$$Q^1_{k+1}(s_k, a_k) = Q^1_k(s_k, a_k) + \alpha \left(r_{k+1} + \gamma Q^2_k(s'_k, \pi_{Q^1_k}(s'_k)) - Q^1_k(s_k, a_k)\right),$$

and $Q^2_{k+1} = Q^2_k$. If $J_k = 2$, the update is defined symmetrically with the roles of Q^1 and Q^2 interchanged. Let $\{\mathcal{F}_k\}_{k\ge0}$ be the natural filtration generated by Q^1_0, Q^2_0 and all samples and table selectors up to time $k - 1$. For $X, Y \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, define the cross-evaluation Bellman map $T_{12}(X, Y) := R + \gamma P\Pi^{\pi_X} Y$. The map T_{21} is defined symmetrically. With these definitions, the exact vector-form of double Q-learning recursion can be expressed as:

$$\begin{aligned} Q^1_{k+1} &= Q^1_k + \alpha p_1 D\big(T_{12}(Q^1_k, Q^2_k) - Q^1_k\big) + \alpha w^1_k, \\ Q^2_{k+1} &= Q^2_k + \alpha p_2 D\big(T_{21}(Q^2_k, Q^1_k) - Q^2_k\big) + \alpha w^2_k, \end{aligned} \tag{1}$$

where the martingale-difference noises are defined for the sampled coordinates as:

$$w^1_k := \mathbf{1}_{\{J_k=1\}}(e_a \otimes e_s)\left(r_{k+1} + \gamma Q^2_k(s'_k, \pi_{Q^1_k}(s'_k)) - Q^1_k(s_k, a_k)\right) - p_1 D\{T_{12}(Q^1_k, Q^2_k) - Q^1_k\},$$

and w^2_k is obtained analogously by exchanging the roles of Q^1_k and Q^2_k. Here, $e_s \in \mathbb{R}^{|\mathcal{S}|}$ and $e_a \in \mathbb{R}^{|\mathcal{A}|}$ denote the standard basis vectors, and $\otimes$ denotes the Kronecker product. Letting $w_k := \begin{bmatrix} w^1_k \\ w^2_k \end{bmatrix}$, it directly follows that $\mathbb{E}[w_k \mid \mathcal{F}_k] = 0$.

4.1 AFFINE SWITCHING-SYSTEM REPRESENTATION OF DOUBLE Q-LEARNING

Next, we apply the affine switching representation to the double Q-learning updates in (1). Define the two-table errors $e^1_k := Q^1_k - Q^*$, and $e^2_k := Q^2_k - Q^*$, and the augmented error vector $x_k := \begin{bmatrix} e^1_k \\ e^2_k \end{bmatrix} \in \mathbb{R}^{2|\mathcal{S}||\mathcal{A}|}$. Let $\pi^1_k := \pi_{Q^1_k}$, and $\pi^2_k := \pi_{Q^2_k}$, and let π^* be an optimal deterministic policy

satisfying $Q^* = R + \gamma P \Pi^{\pi^*} Q^*$. Subtracting this fixed-point equation from the double Q-learning recursion in (1) yields the stochastic affine switching- system:

$$x_{k+1} = A_{\pi_k^1, \pi_k^2} x_k + b_{\pi_k^1, \pi_k^2} + \alpha w_k, \tag{2}$$

where

$$A_{\pi^1,\pi^2} = \begin{bmatrix} I - \alpha p_1 D & \alpha\gamma p_1 DP\Pi^{\pi^1} \\ \alpha\gamma p_2 DP\Pi^{\pi^2} & I - \alpha p_2 D \end{bmatrix}, \; b_{\pi^1,\pi^2} = \alpha\gamma \begin{bmatrix} p_1 DP(\Pi^{\pi^1} - \Pi^{\pi^*})Q^* \\ p_2 DP(\Pi^{\pi^2} - \Pi^{\pi^*})Q^* \end{bmatrix}.$$

Here, $\sigma_k := (\pi_k^1, \pi_k^2)$ denotes the switching mode. The representation is affine because the offset term $b_{\pi_k^1,\pi_k^2}$ is generally nonzero whenever either greedy policy π_k^1 or π_k^2 differs from the optimal policy π^*. This affine offset complicates the direct application of switching-system tools, since the resulting dynamics are no longer purely linear in the error state.

4.2 Direct block switching representation

We now remove the affine offset from the switching-system representation in (2). The key idea is to linearize the cross-evaluation Bellman error through a suitable stochastic-policy representation. The following lemma provides the fundamental identity that allows the affine offset to be absorbed into the switched linear dynamics.

Lemma 2. *For any $X, Y \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$, there exists a stochastic policy μ_X such that*

$$\Pi^{\pi_X} Y - V^* = (\Pi^{\mu_X} - \Pi^{\pi_X})(X - Q^*) + \Pi^{\pi_X}(Y - Q^*).$$

Moreover, μ_X can be chosen measurably as a function of X.

The proof is given in Section B.2 of Appendix. Substituting this identity into the double Q-learning recursion eliminates the affine offset and yields a direct switched-linear representation on the augmented error space, as stated in the following theorem.

Theorem 1. *Under Assumption 1, for each $k \geq 0$, there exist $\mathcal{F}_k$-measurable stochastic policies μ_k^1 and μ_k^2 such that*

$$x_{k+1} = H_k x_k + \alpha w_k, \qquad \mathbb{E}[w_k \mid \mathcal{F}_k] = 0,$$

where

$$H_k = \begin{bmatrix} I - \alpha p_1 D + \alpha\gamma p_1 DP(\Pi^{\mu_k^1} - \Pi^{\pi_k^1}) & \alpha\gamma p_1 DP\Pi^{\pi_k^1} \\ \alpha\gamma p_2 DP\Pi^{\pi_k^2} & I - \alpha p_2 D + \alpha\gamma p_2 DP(\Pi^{\mu_k^2} - \Pi^{\pi_k^2}) \end{bmatrix}.$$

The proof is given in Section B.3 of Appendix.

4.3 Augmented deterministic switching family

The switched-linear representation in Theorem 1 naturally induces an augmented family of deterministic switching modes. We now formalize this family and characterize its stability through the JSR. For $\theta = (\pi_1, \nu_1, \pi_2, \nu_2) \in \Theta^4$, define the extreme augmented mode

$$H_\theta(\alpha) := \begin{bmatrix} I - \alpha p_1 D + \alpha\gamma p_1 DP(\Pi^{\nu_1} - \Pi^{\pi_1}) & \alpha\gamma p_1 DP\Pi^{\pi_1} \\ \alpha\gamma p_2 DP\Pi^{\pi_2} & I - \alpha p_2 D + \alpha\gamma p_2 DP(\Pi^{\nu_2} - \Pi^{\pi_2}) \end{bmatrix}.$$

The augmented double-Q switching family is $\mathcal{H}_\alpha := \{H_\theta(\alpha) : \theta \in \Theta^4\}$. Its JSR is denoted by $\rho_\alpha := \rho(\mathcal{H}_\alpha)$. The following lemma establishes that the conditional-mean matrix H_k in Theorem 1 is generated by the convex hull of the augmented switching family.

Lemma 3. *For every k, the conditional-mean matrix H_k belongs to $\mathrm{co}(\mathcal{H}_\alpha)$. Moreover, $\rho(\mathrm{co}(\mathcal{H}_\alpha)) = \rho(\mathcal{H}_\alpha) = \rho_\alpha$.*

The proof is given in Section B.4 of Appendix.

Remark 1. *In Q-learning, the direct modes are nonnegative and their row sums are easily bounded (Lee, 2026). In double Q-learning, the augmented modes contain signed policy-difference terms $\Pi^{\nu_i} - \Pi^{\pi_i}$. Therefore the nonnegative row-sum argument does not apply to $\mathcal{H}_\alpha$. The condition $\rho_\alpha < 1$ must be verified or replaced by another stability certificate.*

5 MODE-WISE NECESSARY STABILITY CONDITION

While the JSR characterizes worst-case switching behavior, it is often difficult to analyze directly. We therefore derive a necessary step-size condition by examining the eigenvalues of individual switching modes. For each mode $H_\theta(\alpha) \in \mathcal{H}_\alpha$, the dependence on the step-size is affine:

$$H_\theta(\alpha) = I_{2|\mathcal{S}||\mathcal{A}|} + \alpha G_\theta,$$

where

$$G_\theta = \begin{bmatrix} -p_1 D + \gamma p_1 DP(\Pi^{\nu_1} - \Pi^{\pi_1}) & \gamma p_1 DP\Pi^{\pi_1} \\ \gamma p_2 DP\Pi^{\pi_2} & -p_2 D + \gamma p_2 DP(\Pi^{\nu_2} - \Pi^{\pi_2}) \end{bmatrix}.$$

A necessary condition for JSR stability is that every individual mode of the switching family is Schur stable. This observation immediately yields the following mode-wise step-size condition.

Proposition 1. *If $\rho_\alpha < 1$, then, for every $\theta \in \Theta^4$, $\rho(I_{2|\mathcal{S}||\mathcal{A}|} + \alpha G_\theta) < 1$. Equivalently, for every eigenvalue $\lambda \in \operatorname{spec}(G_\theta)$ and every $\theta \in \Theta^4$, $|1 + \alpha\lambda| < 1$. Thus, every such eigenvalue must satisfy $\operatorname{Re}\lambda < 0$, and the step-size must satisfy $0 < \alpha < -2\operatorname{Re}\lambda/|\lambda|^2$ for all $\lambda \in \operatorname{spec}(G_\theta)$ and all $\theta \in \Theta^4$. If some G_θ has an eigenvalue with nonnegative real part, then no positive step-size can make $\rho_\alpha < 1$.*

The proof is given in Section B.5 of Appendix. The following example illustrates the mode-wise instability identified in Proposition 1 and shows that the augmented double-Q switching family need not satisfy $\rho_\alpha < 1$.

Example 1. *Consider a three-state, two-action MDP with*

$$\mathcal{S} = \{1, 2, 3\}, \qquad \mathcal{A} = \{1, 2\}, \qquad \gamma = 0.995.$$

The transitions are deterministic and given by

$$P(3 \mid 1, 1) = 1, \quad P(1 \mid 1, 2) = 1, \quad P(1 \mid 2, 1) = 1,$$
$$P(2 \mid 2, 2) = 1, \quad P(2 \mid 3, 1) = 1, \quad P(1 \mid 3, 2) = 1,$$

with all remaining transition probabilities equal to zero. The rewards are deterministic and given by

$$R(1, 1) = -2, \quad R(1, 2) = -10, \quad R(2, 1) = 6,$$
$$R(2, 2) = 4, \quad R(3, 1) = 10, \quad R(3, 2) = -6.$$

Under the state-action ordering $(1, 1), (2, 1), (3, 1), (1, 2), (2, 2), (3, 2)$, the state-action sampling distribution is

$$D = \operatorname{diag}(0.44,\ 0.0625,\ 0.343,\ 0.0035,\ 0.11,\ 0.041),$$

and the two tables are selected independently with probabilities $p_1 = 0.91,\ p_2 = 0.09$.

Consider the mode $\theta = (\pi_1, \nu_1, \pi_2, \nu_2)$, where $\pi_1 = \pi_2$ and $\nu_1 = \nu_2$ are deterministic policies defined by

$$\big(\pi_1(1), \pi_1(2), \pi_1(3)\big) = (1, 1, 2), \qquad \big(\nu_1(1), \nu_1(2), \nu_1(3)\big) = (2, 2, 1).$$

For this mode, the matrix G_θ has an eigenvalue $\lambda_{\max}(G_\theta) = 0.0914423308 > 0$. Moreover, for every $0 < \alpha < 1$, $\rho(H_\theta(\alpha)) \geq 1 + 0.0914423308\,\alpha > 1$. Consequently, $\rho_\alpha = \rho(\mathcal{H}_\alpha) \geq \rho(H_\theta(\alpha)) > 1$ for every $0 < \alpha < 1$.

Example 1 demonstrates that augmented stability cannot be asserted without a certificate. A corresponding simulation illustrating the larger estimation error of constant step-size double Q-learning compared with standard Q-learning is reported in Section F of Appendix.

6 FINITE-TIME BOUND VIA AN AUGMENTED JSR LYAPUNOV FUNCTION

In this section, we establish a finite-time bound for double Q-learning under an explicit JSR stability condition. Before presenting the main result, we first establish uniform boundedness of the iterates and derive a uniform second-moment bound for the martingale-difference noise. To formalize these bounds, we introduce the following constants: $B_0 := \max\left\{\|Q_0^1\|_\infty, \|Q_0^2\|_\infty, \frac{R_{\max}}{1-\gamma}\right\}$, and $W_{\max} := (R_{\max} + (1+\gamma)B_0)^2$. With these definitions in place, the following lemma shows that the iterates remain uniformly bounded and that the martingale noise has uniformly bounded conditional second moments.

Lemma 4. *Under Assumption 1, the double Q-learning iterates satisfy* $\|Q_k^1\|_\infty \le B_0$, $\|Q_k^2\|_\infty \le B_0$ *and* $\mathbb{E}[\|w_k\|_2^2 \mid \mathcal{F}_k] \le W_{\max}$ *for all* $k \ge 0$.

The proof is given in Section B.6 of Appendix. We next derive a finite-time error bound. The key idea is to compare the switched augmented double-Q recursion with a fixed stochastic reference filter driven by the same martingale noise. Let $\bar{H} \in \mathrm{co}(\mathcal{H}_\alpha)$ be a fixed reference matrix, and consider the reference system

$$\widehat{x}_{k+1} = \bar{H}\widehat{x}_k + \alpha w_k, \qquad \widehat{x}_0 = \begin{bmatrix} Q_0^1 - Q^* \\ Q_0^2 - Q^* \end{bmatrix}.$$

Because the reference system is driven by the same martingale noise as the true recursion, the noise cancels pathwise in the residual recursion. Indeed, with $r_k := x_k - \widehat{x}_k$, we have

$$r_{k+1} = H_k r_k + (H_k - \bar{H})\widehat{x}_k.$$

Thus the martingale noise appears only through the reference filter $\widehat{x}_k$, while the residual recursion is noise-free once $\widehat{x}_k$ and the switching matrices are fixed. To explicitly bound this residual recursion over time, we next establish a geometric product bound whose constants are independent of the step-size. For this purpose, we impose the following fixed-data assumption.

Assumption 2. *The MDP transition kernel, reward function, discount factor, state–action sampling distribution, and table-selection probabilities are fixed independently of* α. *Equivalently,* P, R, γ, D, p_1, p_2 *do not depend on* α.

Under Assumption 2, if $\rho_{\bar{\alpha}} < 1$ for some $\bar{\alpha} > 0$, we establish in Lemma 8 of Appendix that, for all $0 < \alpha \le \bar{\alpha}$, every admissible product of the system matrices admits a uniform geometric bound of the form

$$\|H_{\ell-1} \cdots H_0\|_\infty \le K(1 - c\alpha)^\ell,$$

where K and c are independent of α, yielding the following finite-time bound with an $O(\sqrt{\alpha})$ noise floor.

Theorem 2. *Let Assumption 1 and Assumption 2 hold. Suppose that there exists* $\bar{\alpha} > 0$ *such that* $\rho_{\bar{\alpha}} := \rho_{\mathrm{jsr}}(\mathcal{H}_{\bar{\alpha}}) < 1$. *Let* $c > 0$ *and* $K < \infty$ *be the constants from Lemma 8. Let* $p_{\max} := \max\{p_1, p_2\}$. *For every* $0 < \alpha \le \bar{\alpha}$, *define* $\beta_\alpha := 1 - \frac{c}{2}\alpha$. *Then, for all* $k \ge 0$,

$$\begin{aligned}\mathbb{E}\big[\max\{\|Q_k^1 - Q^*\|_\infty, \|Q_k^2 - Q^*\|_\infty\}\big] \le{}& K\beta_\alpha^k \max\{\|Q_0^1 - Q^*\|_\infty, \|Q_0^2 - Q^*\|_\infty\} \\ &+ 6\gamma p_{\max} d_{\max} \alpha K^2 k \beta_\alpha^{k-1} \\ &\quad \times \max\{\|Q_0^1 - Q^*\|_\infty, \|Q_0^2 - Q^*\|_\infty\} \\ &+ 2|\mathcal{S}||\mathcal{A}|K\left(1 + \frac{6\gamma p_{\max} d_{\max} K}{c}\right)\sqrt{\frac{\alpha W_{\max}}{c}},\end{aligned}$$

where $k\beta_\alpha^{k-1}$ *is interpreted as zero when* $k = 0$. *Consequently,*

$$\limsup_{k\to\infty} \mathbb{E}\big[\max\{\|Q_k^1 - Q^*\|_\infty, \|Q_k^2 - Q^*\|_\infty\}\big] \le 2|\mathcal{S}||\mathcal{A}|K\left(1 + \frac{6\gamma p_{\max} d_{\max} K}{c}\right)\sqrt{\frac{\alpha W_{\max}}{c}}.$$

In particular, the constant-step-size noise floor is $O(\sqrt{\alpha})$. *The same bound also holds for the averaged estimate* $\bar{Q}_k := \frac{Q_k^1 + Q_k^2}{2}$.

The proof is given in Section B.7 of Appendix. The first two terms characterize the transient decay, while the last term represents the constant step-size noise floor.

7 Asymmetric Tracking Double Q-learning

In this section, we propose a novel variant: Asymmetric Tracking double Q-learning (AT-DQ). Unlike standard double Q-learning, this asymmetric architecture is designed to admit a JSR stability certificate under the constant step-size conditions. We first describe the update rule of AT-DQ. At iteration k, the sampled coordinate (s_k, a_k) is updated by

$$Q_{k+1}^1(s_k, a_k) = Q_k^1(s_k, a_k) + \alpha\left(r_{k+1} + \gamma Q_k^2(s_k', \pi_{Q_k^1}(s_k')) - Q_k^1(s_k, a_k)\right),$$

and

$$Q^2_{k+1}(s_k,a_k) = Q^2_k(s_k,a_k) + \beta\left(Q^1_k(s_k,a_k) - Q^2_k(s_k,a_k)\right),$$

with all other coordinates unchanged. Here $\alpha > 0$ is the Bellman-update step-size and $\beta > 0$ is the tracking step-size. The first table selects its greedy action using Q^1_k and evaluates that action using Q^2_k, while the second table tracks the first table directly. The proposed algorithm can be viewed as a tabular analogue of the double DQN target-network architecture in Van Hasselt et al. (2016). The key distinction is that the target table is updated through a soft-tracking recursion with step-size β, rather than by periodic hard copies. This modification enables a direct switching-system analysis while preserving the separation between action selection and action evaluation. To derive its switching-system representation, define $e^1_k := Q^1_k - Q^*$, $e^2_k := Q^2_k - Q^*$, and $x_k := \begin{bmatrix} e^1_k \\ e^2_k \end{bmatrix}$, and let $\pi_k := \pi_{Q^1_k}$. By Lemma 2, there exists an $\mathcal{F}_k$-measurable stochastic policy μ_k such that $\Pi^{\pi_k}Q^2_k - V^* = (\Pi^{\mu_k} - \Pi^{\pi_k})e^1_k + \Pi^{\pi_k}e^2_k$. Therefore the conditional-mean recursion has the form $x_{k+1} = H_k x_k + \zeta_k$, and $\mathbb{E}[\zeta_k \mid \mathcal{F}_k] = 0$, where $H_k = \begin{bmatrix} I - \alpha D + \alpha\gamma DP(\Pi^{\mu_k} - \Pi^{\pi_k}) & \alpha\gamma DP\Pi^{\pi_k} \\ \beta D & I - \beta D \end{bmatrix}$. Consequently, the corresponding deterministic extreme family is $\mathcal{H}^{\mathrm{AT}}_{\alpha,\beta} := \left\{H^{\mathrm{AT}}_{\pi,\nu}(\alpha,\beta) : \pi,\nu \in \Theta\right\}$, where $H_{\pi,\nu}(\alpha,\beta) := \begin{bmatrix} I - \alpha D + \alpha\gamma DP(\Pi^{\nu} - \Pi^{\pi}) & \alpha\gamma DP\Pi^{\pi} \\ \beta D & I - \beta D \end{bmatrix}$. Every H_k belongs to $\mathrm{co}(\mathcal{H}^{\mathrm{AT}}_{\alpha,\beta})$. The present variant contains the signed policy-difference block $\Pi^{\nu} - \Pi^{\pi}$. A direct nonnegative row-sum argument in the original (e^1, e^2) coordinates is therefore not sufficient. The tracking structure, however, yields a simple certificate after the change of variables $\tilde{e}^1_k := Q^1_k - Q^*$ and $\tilde{e}^2_k := Q^2_k - Q^1_k$. For each extreme mode, the transformed deterministic recursion is

$$\tilde{e}^1_{k+1} = (I - \alpha D + \alpha\gamma DP\Pi^{\nu})\tilde{e}^1_k + \alpha\gamma DP\Pi^{\pi}\tilde{e}^2_k,$$

and

$$\tilde{e}^2_{k+1} = \alpha D(I - \gamma P\Pi^{\nu})\tilde{e}^1_k + (I - \beta D - \alpha\gamma DP\Pi^{\pi})\tilde{e}^2_k.$$

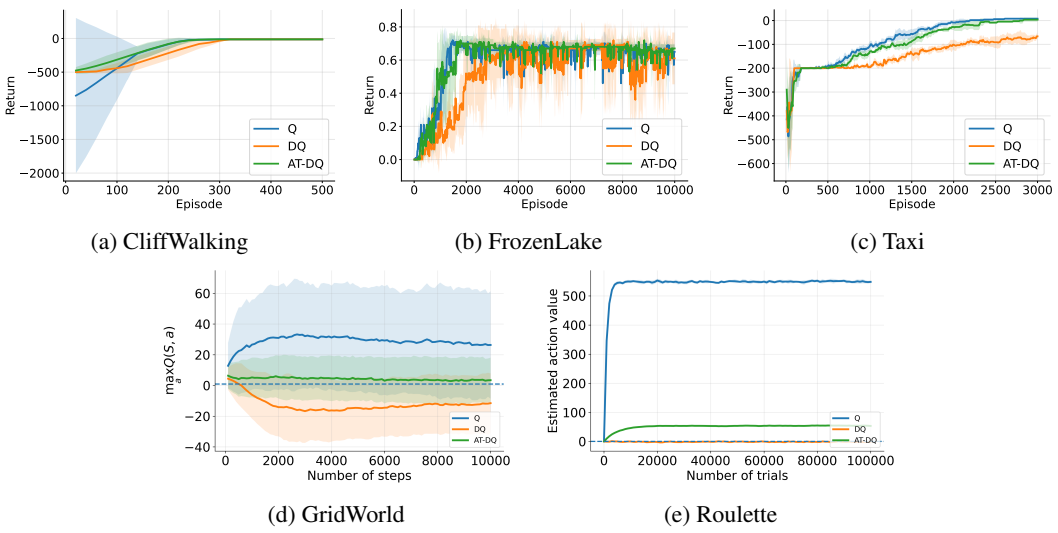
The corresponding deterministic extreme family is $\tilde{\mathcal{H}}^{\mathrm{AT}}_{\alpha,\beta} := \{\tilde{H}^{\mathrm{AT}}_{\pi,\nu}(\alpha,\beta) : \pi,\nu \in \Theta\}$, where $\tilde{H}^{\mathrm{AT}}_{\pi,\nu}(\alpha,\beta) := \begin{bmatrix} I - \alpha D + \alpha\gamma DP\Pi^{\nu} & \alpha\gamma DP\Pi^{\pi} \\ \alpha D(I - \gamma P\Pi^{\nu}) & I - \beta D - \alpha\gamma DP\Pi^{\pi} \end{bmatrix}$. The following proposition shows that the tracking mechanism induces a weighted block-norm contraction, which yields a sufficient step-size condition for JSR stability.

Proposition 2. *Consider the AT-DQ augmented family $\widetilde{\mathcal{H}}^{\mathrm{AT}}_{\alpha,\beta}$. Suppose that the tracking step-size β is proportional to the Bellman-update step-size α, and define $\tau := \frac{\beta}{\alpha}$. Assume that $\tau > \gamma\left(\frac{d_{\max}}{d_{\min}} + \frac{(1+\gamma)d^2_{\max}}{(1-\gamma)d^2_{\min}}\right)$. Choose $\omega \in \left(\frac{\gamma d_{\max}}{d_{\min}(1-\gamma)}, \frac{d_{\min}\tau - \gamma d_{\max}}{(1+\gamma)d_{\max}}\right)$, and define $c := \min\left\{d_{\min}(1-\gamma) - \frac{\gamma d_{\max}}{\omega},\ d_{\min}\tau - \gamma d_{\max} - \omega(1+\gamma)d_{\max}\right\}$. Then $c > 0$. Let $\bar{\alpha} := \min\left\{\frac{1}{d_{\max}}, \frac{1}{\tau d_{\max}}, \frac{1}{c}\right\}$. Then, for every $0 < \alpha \le \bar{\alpha}$ with $\beta = \tau\alpha$, the transformed weighted block sup norm $\left\|\begin{bmatrix}\tilde{e}^1_k \\ \tilde{e}^2_k\end{bmatrix}\right\|_\omega := \max\left\{\|\tilde{e}^1_k\|_\infty, \omega\|\tilde{e}^2_k\|_\infty\right\}$ satisfies $\left\|\begin{bmatrix}\tilde{e}^1_{k+1} \\ \tilde{e}^2_{k+1}\end{bmatrix}\right\|_\omega \le (1-\alpha c)\left\|\begin{bmatrix}\tilde{e}^1_k \\ \tilde{e}^2_k\end{bmatrix}\right\|_\omega$ for every extreme mode of the AT-DQ switching family. The same contraction bound holds for all matrices in $\mathrm{co}(\widetilde{\mathcal{H}}^{\mathrm{AT}}_{\alpha,\beta})$. Consequently, $\rho_{\mathrm{jsr}}\left(\mathrm{co}(\widetilde{\mathcal{H}}^{\mathrm{AT}}_{\alpha,\beta})\right) = \rho_{\mathrm{jsr}}\left(\widetilde{\mathcal{H}}^{\mathrm{AT}}_{\alpha,\beta}\right) \le 1 - \alpha c < 1$.*

The proof is given in Section B.8 of Appendix. A closely related greedy-selection tracking variant can also be constructed by letting the tracking table select the greedy action while the primary table performs the Bellman evaluation. This variant also admits an analogous JSR stability certificate and is presented in Section D of Appendix.

8 Experiments

In this section, we empirically evaluate the performance of AT-DQ against standard baselines. We conduct experiments on classic control tasks from OpenAI Gym (Brockman et al., 2016), as well as on Roulette and Gridworld in Hasselt (2010), which are well-known environments designed to illustrate the maximization bias. For brevity, detailed environment descriptions and hyperparameter configurations are deferred to Section G of Appendix.



(a) CliffWalking (b) FrozenLake (c) Taxi

(d) GridWorld (e) Roulette

Figure 1: Performance comparison of Q-learning, double Q-learning, and AT-DQ. (a)–(c) show convergence on CliffWalking, FrozenLake, and Taxi, while (d)–(e) show maximization bias behavior on GridWorld and Roulette.

8.1 Convergence

To demonstrate the convergence properties of our proposed algorithm, we evaluate it in standard Gym environments, specifically FrozenLake, CliffWalking, and Taxi. We use $\alpha = 0.1$ for all algorithms and $\beta = 0.8$ for AT-DQ. All reported results are averaged over 10 random seeds. As illustrated in Figure 1(a)–Figure 1(c), AT-DQ demonstrates superior convergence performance across all evaluated environments compared to double Q-learning. We also provide an ablation study on the ratio β/α in Section H.1 of Appendix.

8.2 Maximization Bias

To investigate the effect of maximization bias, we conduct experiments on the environments Roulette and Gridworld. For all algorithms, we use $\alpha = 0.1$, while AT-DQ employs a tracking step-size of $\beta = 0.025$. The results are averaged over 10 random seeds in Roulette and 1000 random seeds in Gridworld. The dotted horizontal line indicates the true optimal start-state value in Gridworld and the true value of the termination action in Roulette. In the Roulette environment, our proposed algorithm exhibits a slight overestimation. While it does not eliminate the bias as effectively as double Q-learning, it still achieves a meaningful reduction in maximization bias compared to standard Q-learning. Conversely, in the Gridworld environment, AT-DQ successfully mitigates the maximization bias inherent in Q-learning, whereas double Q-learning suffers from underestimation. We further investigate the effect of the ratio β/α on the maximization bias of AT-DQ, and defer the detailed results to Section H.2 of Appendix.

9 Conclusion

This paper developed a direct switching-system analysis for double Q-learning. The resulting augmented switching family contains signed policy-difference blocks, showing that JSR stability is not automatically guaranteed. We therefore established worst-case finite-time bounds conditionally on an explicit JSR stability condition. To overcome this limitation, we proposed an asymmetric variant, AT-DQ, that is provably JSR-stable under constant step sizes and exhibits competitive empirical performance compared with standard double Q-learning.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of the manuscript, to assist with retrieval and discovery of relevant literature, and to assist in an exploratory search for an example containing an unstable switching mode presented in Example 1. The resulting example and all associated numerical and theoretical claims were independently verified by the authors. We reviewed all AI-assisted work and take responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

This work primarily makes theoretical contributions. The assumptions, the direct switching-system formulation of double Q-learning, the JSR-based stability analysis, and the proposed AT-DQ algorithm are stated in the main text. The JSR-induced common Lyapunov construction, complete proofs, and supporting finite-time analyses are provided in Sections A to C of Appendix. The explicit mode calculation for the unstable switching-mode counterexample and the corresponding sample-path simulation are provided in Sections E and F of Appendix. Detailed environment descriptions and hyperparameter settings for the experiments are provided in Section G of Appendix, and additional ablation studies on the tracking ratio β/α are reported in Sections H.1 and H.2 of Appendix.

REFERENCES

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Heng Du, Ligang Cong, Qingyun Liang, and Shuyu Wang. Sdrp-dql: A satellite distributed routing protocol based on double deep q-learning. *Computer Networks*, pp. 112068, 2026.

Xiaoliang Fan, Chunguang Bu, Xingang Zhao, Jin Sui, and Hongwei Mo. Incremental double q-learning enhanced mpc for trajectory tracking of mobile robots. *IEEE Transactions on Instrumentation and Measurement*, 2025.

Hado V Hasselt. Double Q-learning. In *Advances in Neural Information Processing Systems*, pp. 2613–2621, 2010.

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

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

Donghwan Lee. Lyapunov-certified direct switching theory for q-learning. *arXiv preprint arXiv:2604.19569*, 2026.

Donghwan Lee and Han-Dong Lim. A switching system theory of q-learning with linear function approximation. *arXiv preprint arXiv:2605.11021*, 2026.

Daniel Liberzon. *Switching in systems and control*. Springer Science & Business Media, 2003.

Hyunjun Na and Donghwan Lee. Finite-time analysis of simultaneous double q-learning. *Neurocomputing*, pp. 132581, 2026.

Mayur Patil, Pooja Tambolkar, and Shawn Midlam-Mohler. Optimizing traffic routes with enhanced double q-learning. *IET Intelligent Transport Systems*, 19(1):e70002, 2025.

Martin L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Gian-Carlo Rota and Gilbert Strang. A note on the joint spectral radius. *Indag. Math*, 22(4):379–381, 1960.

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

John N Tsitsiklis and Vincent D Blondel. The lyapunov exponent and joint spectral radius of pairs of matrices are hardwhen not impossibleto compute and to approximate. *Mathematics of Control, Signals and Systems*, 10(1):31–40, 1997.

Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

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

Huaqing Xiong, Lin Zhao, Yingbin Liang, and Wei Zhang. Finite-time analysis for double q-learning. *Advances in neural information processing systems*, 33:16628–16638, 2020.

Mohamed N Zareer and Rastko R Selmic. Maximizing opinion polarization using double deep q-learning in social networks. *IEEE Access*, 13:57398–57412, 2025.

Lin Zhao, Huaqing Xiong, and Yingbin Liang. Faster non-asymptotic convergence for double q-learning. *Advances in Neural Information Processing Systems*, 34:7242–7253, 2021.

A JSR-INDUCED COMMON LYAPUNOV NORM

This section presents a standard finite-family JSR construction used throughout the subsequent analysis. In particular, it converts the strict JSR condition into a common Lyapunov norm that contracts every matrix in the switching family. We include the construction to make the later stability and finite-time arguments self-contained.

Lemma 5. *(Lee & Lim, 2026) Let*

$$\mathcal{H} = \{H_1, H_2, \dots, H_M\} \subset \mathbb{R}^{d\times d}, \qquad \rho := \rho_{\mathrm{jsr}}(\mathcal{H}),$$

and fix $\varepsilon > 0$ such that

$$\beta_\varepsilon := \rho + \varepsilon \in (0, 1).$$

For a mode sequence

$$\sigma = (\sigma_1, \dots, \sigma_k) \in \{1, \dots, M\}^k,$$

define

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

with the convention that $H_\sigma = I$ when $k = 0$. For every integer $t \geq 0$, define

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

Then the following statements hold.

(i) For every $t \geq 0$,

$$V_\varepsilon^{t+1}(x) \geq \|x\|_2^2 + \beta_\varepsilon^{-2} \max_{i\in\{1,\dots,M\}} V_\varepsilon^t(H_i x), \qquad \forall x \in \mathbb{R}^d.$$

(ii) For every $t \geq 0$, the function V_ε^t is absolutely homogeneous of degree two and is nondecreasing in t:

$$V_\varepsilon^t(\lambda x) = |\lambda|^2 V_\varepsilon^t(x), \qquad \forall \lambda \in \mathbb{R}, \quad \forall x \in \mathbb{R}^d,$$

and

$$V_\varepsilon^t(x) \leq V_\varepsilon^{t+1}(x), \qquad \forall x \in \mathbb{R}^d.$$

(iii) There exists a constant $C_\varepsilon > 0$ such that

$$\|x\|_2^2 \leq V_\varepsilon^t(x) \leq C_\varepsilon \|x\|_2^2, \qquad \forall x \in \mathbb{R}^d, \quad \forall t \geq 0.$$

(iv) The pointwise limit

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

exists and satisfies

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

(v) The function

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

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

(vi) For every $i \in \{1, \dots, M\}$,

$$V_\varepsilon^\infty(H_i x) \leq \beta_\varepsilon^2 \left(V_\varepsilon^\infty(x) - \|x\|_2^2\right) \leq \beta_\varepsilon^2 V_\varepsilon^\infty(x), \qquad \forall x \in \mathbb{R}^d.$$

Equivalently,

$$p_\varepsilon(H_i x) \leq \beta_\varepsilon p_\varepsilon(x), \qquad \forall x \in \mathbb{R}^d.$$

The resulting norm p_ε provides a common contraction certificate for all deterministic modes in $\mathcal{H}_\alpha$.

B PROOF

B.1 PROOF OF LEMMA 1

Lemma 1 (restated). For every $Q \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$, there exists a stochastic policy $\mu_Q : \mathcal{S} \to \Delta_{|\mathcal{A}|}$ such that

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

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

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

$$e := Q - Q^*.$$

Let

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

Since

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

we have

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

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

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

so

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

Thus y_s belongs to the convex hull of the finite set $\{e(s,a) : a \in \mathcal{A}\}$. Therefore it can be written as a convex combination of the action-wise errors at state s.

Choose such a convex combination for every state. Stacking the resulting state-wise identities gives

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

A measurable selection is obtained by lexicographically breaking ties among minimizers and maximizers and choosing the corresponding convex coefficient on each finite comparison region. □

B.2 PROOF OF LEMMA 2

Lemma 2 (restated). For any $X, Y \in \mathbb{R}^{|\mathcal{S}|\,|\mathcal{A}|}$, there exists a stochastic policy μ_X such that

$$\Pi^{\pi_X} Y - V^* = (\Pi^{\mu_X} - \Pi^{\pi_X})(X - Q^*) + \Pi^{\pi_X}(Y - Q^*).$$

Moreover, μ_X can be chosen measurably as a function of X.

Proof. Apply Lemma 1 to the table X. There exists a stochastic policy μ_X such that

$$V_X - V^* = \Pi^{\mu_X}(X - Q^*).$$

Because π_X is greedy with respect to X,

$$\Pi^{\pi_X} X = V_X.$$

Now decompose

$$\begin{aligned}
\Pi^{\pi_X} Y - V^* &= \Pi^{\pi_X}(Y - X) + \Pi^{\pi_X} X - V^* \\
&= \Pi^{\pi_X}(Y - X) + V_X - V^* \\
&= \Pi^{\pi_X}(Y - X) + \Pi^{\mu_X}(X - Q^*) \\
&= \Pi^{\pi_X}(Y - Q^*) - \Pi^{\pi_X}(X - Q^*) + \Pi^{\mu_X}(X - Q^*) \\
&= (\Pi^{\mu_X} - \Pi^{\pi_X})(X - Q^*) + \Pi^{\pi_X}(Y - Q^*).
\end{aligned}$$

Measurability follows from the measurable selection in Lemma 1. □

B.3 Proof of Theorem 1

Theorem 1 (restated). Under Assumption 1, for each $k \geq 0$, there exist $\mathcal{F}_k$-measurable stochastic policies μ_k^1 and μ_k^2 such that

$$x_{k+1} = H_k x_k + \alpha w_k, \qquad \mathbb{E}[w_k \mid \mathcal{F}_k] = 0,$$

where

$$H_k = \begin{bmatrix} I - \alpha p_1 D + \alpha\gamma p_1 DP(\Pi^{\mu_k^1} - \Pi^{\pi_k^1}) & \alpha\gamma p_1 DP\Pi^{\pi_k^1} \\ \alpha\gamma p_2 DP\Pi^{\pi_k^2} & I - \alpha p_2 D + \alpha\gamma p_2 DP(\Pi^{\mu_k^2} - \Pi^{\pi_k^2}) \end{bmatrix}.$$

Proof. The first table satisfies

$$Q_{k+1}^1 = Q_k^1 + \alpha p_1 D\{T_{12}(Q_k^1, Q_k^2) - Q_k^1\} + \alpha w_k^1.$$

Subtract Q^* from both sides. Since $Q^* = R + \gamma P V^*$,

$$T_{12}(Q_k^1, Q_k^2) - Q^* = \gamma P(\Pi^{\pi_k^1} Q_k^2 - V^*).$$

By Lemma 2 with $X = Q_k^1$ and $Y = Q_k^2$, there exists an $\mathcal{F}_k$-measurable stochastic policy μ_k^1 such that

$$\Pi^{\pi_k^1} Q_k^2 - V^* = (\Pi^{\mu_k^1} - \Pi^{\pi_k^1}) e_k^1 + \Pi^{\pi_k^1} e_k^2.$$

Therefore

$$\begin{aligned} e_{k+1}^1 &= e_k^1 + \alpha p_1 D\{T_{12}(Q_k^1, Q_k^2) - Q^* - e_k^1\} + \alpha w_k^1 \\ &= \left(I - \alpha p_1 D + \alpha\gamma p_1 DP(\Pi^{\mu_k^1} - \Pi^{\pi_k^1})\right) e_k^1 + \alpha\gamma p_1 DP\Pi^{\pi_k^1} e_k^2 + \alpha w_k^1. \end{aligned}$$

The same argument with the two tables exchanged gives an $\mathcal{F}_k$-measurable stochastic policy μ_k^2 satisfying

$$e_{k+1}^2 = \alpha\gamma p_2 DP\Pi^{\pi_k^2} e_k^1 + \left(I - \alpha p_2 D + \alpha\gamma p_2 DP(\Pi^{\mu_k^2} - \Pi^{\pi_k^2})\right) e_k^2 + \alpha w_k^2.$$

Stacking the two equations proves the result. □

B.4 Proof of Lemma 3

Lemma 3 (restated). For every k, the conditional-mean matrix H_k belongs to $\mathrm{co}(\mathcal{H}_\alpha)$. Moreover,

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

Proof. Every stochastic policy is a convex combination of deterministic policies. Indeed, for a stochastic policy μ, define

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

Then $c_\pi(\mu) \geq 0$ and $\sum_{\pi \in \Theta} c_\pi(\mu) = 1$, and

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

Applying this representation to μ_k^1 and μ_k^2 shows that H_k is a convex combination of the deterministic extreme modes in $\mathcal{H}_\alpha$.

The JSR convex-hull equality follows from the standard argument: a product of convex combinations expands into a convex combination of products of extreme matrices. Convexity of any norm gives the upper bound

$$\rho(\mathrm{co}(\mathcal{H}_\alpha)) \leq \rho(\mathcal{H}_\alpha),$$

while the reverse inequality follows from set inclusion. □

B.5 Proof of Proposition 1

Proposition 1 (restated). If $\rho_\alpha < 1$, then, for every $\theta \in \Theta^4$, $\rho(I_{2|\mathcal{S}||\mathcal{A}|} + \alpha G_\theta) < 1$. Equivalently, for every eigenvalue $\lambda \in \mathrm{spec}(G_\theta)$ and every $\theta \in \Theta^4$, $|1 + \alpha\lambda| < 1$. Thus, every such eigenvalue must satisfy $\mathrm{Re}\,\lambda < 0$, and the step-size must satisfy $0 < \alpha < -2\,\mathrm{Re}\,\lambda/|\lambda|^2$ for all $\lambda \in \mathrm{spec}(G_\theta)$ and all $\theta \in \Theta^4$. If some G_θ has an eigenvalue with nonnegative real part, then no positive step-size can make $\rho_\alpha < 1$.

Proof. The JSR of a family is at least the spectral radius of each individual member. Thus $\rho_\alpha < 1$ implies

$$\rho(H_\theta(\alpha)) < 1$$

for every θ. Since

$$H_\theta(\alpha) = I + \alpha G_\theta,$$

the eigenvalues of $H_\theta(\alpha)$ are $1 + \alpha\lambda$, where λ ranges over the eigenvalues of G_θ. Therefore Schur stability of each mode requires

$$|1 + \alpha\lambda| < 1.$$

Expanding the squared modulus gives

$$|1 + \alpha\lambda|^2 = 1 + 2\alpha\,\mathrm{Re}\,\lambda + \alpha^2|\lambda|^2.$$

The condition $|1 + \alpha\lambda| < 1$ is equivalent to

$$2\,\mathrm{Re}\,\lambda + \alpha|\lambda|^2 < 0.$$

For $\alpha > 0$, this can hold only if $\mathrm{Re}\,\lambda < 0$, and then it is equivalent to

$$0 < \alpha < \frac{-2\,\mathrm{Re}\,\lambda}{|\lambda|^2}.$$

□

B.6 Proof of Lemma 4

Lemma 4 (restated). Under Assumption 1, the double Q-learning iterates satisfy

$$\|Q_k^1\|_\infty \le B_0, \qquad \|Q_k^2\|_\infty \le B_0, \qquad \forall k \ge 0.$$

Moreover,

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

Proof. We prove that $\|Q_k^1\|_\infty \le B_0$ and $\|Q_k^2\|_\infty \le B_0$ for all $k \ge 0$ by mathematical induction on k, adopting an asynchronous update structure.

At $k = 0$, the bounds $\|Q_0^1\|_\infty \le B_0$ and $\|Q_0^2\|_\infty \le B_0$ trivially hold by the definition of B_0. Assume that for some iteration m, the bounds $|Q_m^1(s,a)| \le B_0$ and $|Q_m^2(s,a)| \le B_0$ hold for all state-action pairs (s,a).

At the $(m+1)$-th iteration, the update is performed for a specific state-action pair (s,a) on only one of the tables. If Table 1 is updated, applying the triangle inequality to the update equation yields

$$|Q_{m+1}^1(s,a)| \le (1-\alpha)|Q_m^1(s,a)| + \alpha\big(|r_{m+1}| + \gamma|Q_m^2(s', \pi_{Q_m^1}(s'))|\big).$$

Applying the inductive hypothesis and the definition of $R_{\max}$, this value is bounded by $(1-\alpha)B_0 + \alpha(R_{\max} + \gamma B_0)$. Since $B_0 \ge R_{\max}/(1-\gamma)$ by definition, it follows that $R_{\max} \le (1-\gamma)B_0$. Substituting this inequality simplifies the bound to exactly B_0. By symmetry, the same bound holds if Table 2 is updated. For all coordinates not selected for an update, the values remain unchanged and thus are still bounded by B_0. Therefore, by mathematical induction, the iterates of both tables are uniformly bounded by B_0 for all $k \ge 0$.

To establish the noise bound, let

$$Z_k := \begin{bmatrix} Z_k^1 \\ Z_k^2 \end{bmatrix}$$

denote the uncentered stochastic update vector at iteration k before scaling by the step-size α, where

$$Z_k^1 = \mathbf{1}_{\{J_k=1\}}(e_{a_k} \otimes e_{s_k}) \left(r_{k+1} + \gamma Q_k^2\big(s'_k, \pi_{Q_k^1}(s'_k)\big) - Q_k^1(s_k, a_k) \right),$$

and

$$Z_k^2 = \mathbf{1}_{\{J_k=2\}}(e_{a_k} \otimes e_{s_k}) \left(r_{k+1} + \gamma Q_k^1\big(s'_k, \pi_{Q_k^2}(s'_k)\big) - Q_k^2(s_k, a_k) \right).$$

At each iteration, this vector has at most one nonzero coordinate. If the update occurs for Table 1, the absolute value of this target coordinate is bounded by $|r_{k+1}| + \gamma|Q_k^2(s'_k, \pi_{Q_k^1}(s'_k))| + |Q_k^1(s_k, a_k)|$. Using the uniform bound established above, this does not exceed $R_{\max} + (1+\gamma)B_0$. Because only one coordinate is updated across both tables, its squared Euclidean norm satisfies

$$\|Z_k\|_2^2 \le (R_{\max} + (1+\gamma)B_0)^2 = W_{\max}.$$

The martingale difference noise is defined as the conditionally centered update $w_k = Z_k - \mathbb{E}[Z_k \mid \mathcal{F}_k]$. Therefore, the conditional second moment satisfies

$$\begin{aligned} \mathbb{E}\big[\|w_k\|_2^2 \mid \mathcal{F}_k\big] &= \mathbb{E}\big[\|Z_k\|_2^2 \mid \mathcal{F}_k\big] - \|\mathbb{E}[Z_k \mid \mathcal{F}_k]\|_2^2 \\ &\le \mathbb{E}\big[\|Z_k\|_2^2 \mid \mathcal{F}_k\big] \le W_{\max}. \end{aligned}$$

□

B.7 Proof of Theorem 2

Before proving the theorem, we record a standard consequence of the JSR condition. It converts the strict JSR inequality $\rho_\alpha < 1$ into a uniform product-growth bound for all matrix products generated by $\mathrm{co}(\mathcal{H}_\alpha)$.

Lemma 6. *(Lee, 2026) Fix any ϵ such that $\beta_\varepsilon := \rho_\alpha + \varepsilon < 1$. Define*

$$K_{\beta_\epsilon} := \sup_{\ell \geq 0} \beta_\epsilon^{-\ell} \sup_{H_0,\dots,H_{\ell-1} \in \mathrm{co}(\mathcal{H}_\alpha)} \|H_{\ell-1} \cdots H_0\|_\infty ,$$

where the product of length zero is interpreted as the identity. Then $K_{\beta_\epsilon} < \infty$. Moreover, for every $\ell \geq 0$ and every sequence $H_0, \dots, H_{\ell-1} \in \mathrm{co}(\mathcal{H}_\alpha)$,

$$\|H_{\ell-1} \cdots H_0\|_\infty \leq K_{\beta_\epsilon} \beta_\epsilon^\ell.$$

Proof. By Lemma 3,

$$\rho_{\mathrm{jsr}}(\mathrm{co}(\mathcal{H}_\alpha)) = \rho_{\mathrm{jsr}}(\mathcal{H}_\alpha) = \rho_\alpha.$$

Since $\beta_\epsilon > \rho_\alpha$, the definition of the JSR implies that products generated by $\mathrm{co}(\mathcal{H}_\alpha)$ grow at a rate strictly smaller than β_ϵ. Equivalently,

$$\sup_{\ell \geq 0} \beta_\epsilon^{-\ell} \sup_{H_0,\dots,H_{\ell-1} \in \mathrm{co}(\mathcal{H}_\alpha)} \|H_{\ell-1} \cdots H_0\|_\infty < \infty.$$

Therefore $K_{\beta_\epsilon} < \infty$. The product bound follows directly from the definition of K_{β_ϵ}:

$$\|H_{\ell-1} \cdots H_0\|_\infty \leq K_{\beta_\epsilon} \beta_\epsilon^\ell.$$

This proves the lemma. □

Although Lemma 6 guarantees that $K_{\beta_\epsilon} < \infty$ for each fixed step-size α, the constant K_{β_ϵ} may itself depend on α. We next show that, under Assumption 2 and JSR stability at one reference step-size, the product-growth bound can be chosen with a multiplicative constant that is uniform over all smaller step sizes.

To establish this uniformity, we proceed in two steps. First, we exploit the affine dependence of the deterministic modes on α to transfer JSR stability at the reference step-size to a common one-step contraction of the form $1 - c\alpha$ for every smaller step-size. Second, we convert this uniform contraction into an infinity-norm product bound whose multiplicative constant is independent of α.

The first step relies on the affine step-size dependence implied by Assumption 2. In particular, for

$$\theta = (\pi_1, \nu_1, \pi_2, \nu_2) \in \Theta^4,$$

we have

$$H_\theta(\alpha) = I_{2|\mathcal{S}||\mathcal{A}|} + \alpha G_\theta,$$

The following lemma carries out the first step: stability at one reference step-size yields a common linear contraction gap for every smaller step-size.

Lemma 7. *(Lee & Lim, 2026) Suppose that Assumption 2 holds and that there exists $\bar{\alpha} > 0$ such that*

$$\rho_{\bar{\alpha}} := \rho_{\mathrm{jsr}}(\mathcal{H}_{\bar{\alpha}}) < 1.$$

Then there exist a norm p on $\mathbb{R}^{2|\mathcal{S}||\mathcal{A}|}$ and a constant $c > 0$, both independent of α, such that, for every $0 < \alpha \leq \bar{\alpha}$,

$$p(Hx) \leq (1 - c\alpha)p(x), \qquad \forall x \in \mathbb{R}^{2|\mathcal{S}||\mathcal{A}|}, \quad \forall H \in \mathrm{co}(\mathcal{H}_\alpha).$$

Proof. Choose η such that

$$\rho_{\bar{\alpha}} < \eta < 1.$$

By the common product-norm construction in Lemma 5, applied to the finite family $\mathcal{H}_{\bar{\alpha}}$, there exists a norm p such that

$$p\big(H_\theta(\bar{\alpha})x\big) \leq \eta p(x), \qquad \forall x \in \mathbb{R}^{2|\mathcal{S}||\mathcal{A}|}, \quad \forall \theta \in \Theta^4.$$

By convexity of the norm, the same inequality holds for every $B \in \mathrm{co}(\mathcal{H}_{\bar{\alpha}})$:

$$p(Bx) \le \eta p(x), \qquad \forall x \in \mathbb{R}^{2|\mathcal{S}||\mathcal{A}|}.$$

Fix $0 < \alpha \le \bar{\alpha}$, and define

$$t := \frac{\alpha}{\bar{\alpha}} \in (0, 1].$$

For every $\theta \in \Theta^4$,

$$\begin{aligned} H_\theta(\alpha) &= I_{2|\mathcal{S}||\mathcal{A}|} + \alpha G_\theta \\ &= I_{2|\mathcal{S}||\mathcal{A}|} + t\bar{\alpha} G_\theta \\ &= (1-t) I_{2|\mathcal{S}||\mathcal{A}|} + t H_\theta(\bar{\alpha}). \end{aligned}$$

Consequently, every $H \in \mathrm{co}(\mathcal{H}_\alpha)$ can be written as

$$H = (1-t) I_{2|\mathcal{S}||\mathcal{A}|} + tB$$

for some $B \in \mathrm{co}(\mathcal{H}_{\bar{\alpha}})$. Therefore,

$$\begin{aligned} p(Hx) &\le (1-t)p(x) + tp(Bx) \\ &\le \big((1-t) + t\eta\big) p(x) \\ &= \left(1 - \frac{1-\eta}{\bar{\alpha}}\alpha\right) p(x). \end{aligned}$$

Defining

$$c := \frac{1-\eta}{\bar{\alpha}} > 0$$

proves the result. □

Lemma 7 provides a common contraction norm for all $0 < \alpha \le \bar{\alpha}$. The next lemma converts this contraction into an infinity-norm product bound whose constants are uniform in the step-size.

Lemma 8. *(Lee & Lim, 2026) Suppose that Assumption 2 holds and that there exists $\bar{\alpha} > 0$ such that*

$$\rho_{\bar{\alpha}} < 1.$$

Then there exist constants $c > 0$ and $K < \infty$, independent of α, such that, for every $0 < \alpha \le \bar{\alpha}$,

$$\rho_\alpha \le 1 - c\alpha < 1,$$

and

$$\sup_{H_0, \dots, H_{\ell-1} \in \mathrm{co}(\mathcal{H}_\alpha)} \|H_{\ell-1} \cdots H_0\|_\infty \le K(1-c\alpha)^\ell, \qquad \forall \ell \ge 0. \tag{3}$$

Proof. Let p and c be given by Lemma 7. Repeated application of the one-step inequality gives

$$p(H_{\ell-1} \cdots H_0 x) \le (1-c\alpha)^\ell p(x)$$

for every $H_0, \dots, H_{\ell-1} \in \mathrm{co}(\mathcal{H}_\alpha)$.

Taking ℓ-th roots, maximizing over all such products, and passing to the limit in the definition of the JSR gives

$$\rho_{\mathrm{jsr}}\big(\mathrm{co}(\mathcal{H}_\alpha)\big) \le 1 - c\alpha.$$

By the convex-hull invariance established in Lemma 3,

$$\rho_\alpha = \rho_{\mathrm{jsr}}(\mathcal{H}_\alpha) = \rho_{\mathrm{jsr}}\big(\mathrm{co}(\mathcal{H}_\alpha)\big) \le 1 - c\alpha.$$

Moreover, the construction in Lemma 7 gives

$$1 - c\alpha = 1 - (1-\eta)\frac{\alpha}{\bar{\alpha}} \ge \eta > 0$$

for every $0 < \alpha \le \bar{\alpha}$. Hence

$$0 < 1 - c\alpha < 1.$$

Since all norms on $\mathbb{R}^{2|\mathcal{S}||\mathcal{A}|}$ are equivalent, there exist constants $a, b < \infty$, independent of α, such that

$$\|x\|_\infty \le a p(x), \qquad p(x) \le b\|x\|_\infty, \qquad \forall x \in \mathbb{R}^{2|\mathcal{S}||\mathcal{A}|}.$$

Therefore,

$$\begin{aligned}\|H_{\ell-1}\cdots H_0 x\|_\infty &\le a\, p(H_{\ell-1}\cdots H_0 x)\\ &\le a(1-c\alpha)^\ell p(x)\\ &\le ab(1-c\alpha)^\ell \|x\|_\infty.\end{aligned}$$

Setting $K := ab$ and taking the induced infinity norm proves (3). □

The next lemma shows that the product-growth constant introduced in Lemma 6 can be chosen uniformly over all $0 < \alpha \le \bar{\alpha}$.

Lemma 9. *(Lee & Lim, 2026) Under the assumptions of Lemma 8, let $c > 0$ and $K < \infty$ be the constants in (3). Define*

$$\beta_\alpha := 1 - \frac{c}{2}\alpha.$$

Then

$$\rho_\alpha < \beta_\alpha < 1,$$

and the product-growth constant in Lemma 6 satisfies

$$K_{\beta_\alpha} \le K, \qquad 0 < \alpha \le \bar{\alpha}.$$

Proof. By Lemma 8,

$$\rho_\alpha \le 1 - c\alpha < 1 - \frac{c}{2}\alpha = \beta_\alpha.$$

Moreover, since $c\alpha < 1$, we also have

$$0 < \beta_\alpha < 1.$$

Using (3) in the definition of K_{β_α} gives

$$\begin{aligned}K_{\beta_\alpha} &= \sup_{\ell\ge 0}\ \sup_{H_0,\dots,H_{\ell-1}\in \mathrm{co}(\mathcal{H}_\alpha)} \beta_\alpha^{-\ell}\,\|H_{\ell-1}\cdots H_0\|_\infty\\ &\le K \sup_{\ell\ge 0}\left(\frac{1-c\alpha}{\beta_\alpha}\right)^\ell\\ &= K \sup_{\ell\ge 0}\left(\frac{1-c\alpha}{1-\frac{c}{2}\alpha}\right)^\ell.\end{aligned}$$

Because

$$0 \le \frac{1-c\alpha}{1-\frac{c}{2}\alpha} < 1,$$

the supremum is attained at $\ell = 0$. Consequently,

$$K_{\beta_\alpha} \le K.$$

□

Theorem 2 (restated). Let Assumption 1 and Assumption 2 hold. Suppose that there exists $\bar{\alpha} > 0$ such that $\rho_{\bar{\alpha}} := \rho_{\mathrm{jsr}}(\mathcal{H}_{\bar{\alpha}}) < 1$. Let $c > 0$ and $K < \infty$ be the constants from Lemma 8. Let $p_{\max} := \max\{p_1, p_2\}$. For every $0 < \alpha \le \bar{\alpha}$, define $\beta_\alpha := 1 - \frac{c}{2}\alpha$. Then, for all $k \ge 0$,

$$\begin{aligned}\mathbb{E}\big[\max\{\|Q_k^1 - Q^*\|_\infty, \|Q_k^2 - Q^*\|_\infty\}\big] &\le K\beta_\alpha^k \max\{\|Q_0^1 - Q^*\|_\infty, \|Q_0^2 - Q^*\|_\infty\}_\infty \\ &\quad + 6\gamma p_{\max} d_{\max} \alpha K^2 k \beta_\alpha^{k-1} \\ &\quad \times \max\{\|Q_0^1 - Q^*\|_\infty, \|Q_0^2 - Q^*\|_\infty\} \\ &\quad + 2|\mathcal{S}||\mathcal{A}|K\left(1 + \frac{6\gamma p_{\max} d_{\max} K}{c}\right)\sqrt{\frac{\alpha W_{\max}}{c}},\end{aligned}$$

where $k\beta_\alpha^{k-1}$ is interpreted as zero when $k = 0$. Consequently,

$$\begin{aligned}\limsup_{k\to\infty} \mathbb{E}\big[\max\{\|Q_k^1 - Q^*\|_\infty, \|Q_k^2 - Q^*\|_\infty\}\big] &\le 2|\mathcal{S}||\mathcal{A}|K\left(1 + \frac{6\gamma p_{\max} d_{\max} K}{c}\right) \\ &\quad \times \sqrt{\frac{\alpha W_{\max}}{c}}.\end{aligned}$$

In particular, the constant-step-size noise floor is $O(\sqrt{\alpha})$. The same bound also holds for the averaged estimate $\bar{Q}_k := \frac{Q_k^1 + Q_k^2}{2}$.

Proof. Fix $0 < \alpha \le \bar{\alpha}$, by Lemma 8,

$$0 < 1 - c\alpha < \beta_\alpha < 1,$$

and

$$\sup_{H_0,\dots,H_{\ell-1} \in \mathrm{co}(\mathcal{H}_\alpha)} \|H_{\ell-1} \cdots H_0\|_\infty \le K(1 - c\alpha)^\ell, \qquad \forall \ell \ge 0. \tag{4}$$

We first establish two consequences of (4) that will be used in the subsequent analysis. For every fixed $\bar{H} \in \mathrm{co}(\mathcal{H}_\alpha)$,

$$\begin{aligned}\sum_{i=0}^{k-1} \|\bar{H}^i\|_\infty^2 &\le K^2 \sum_{i=0}^{k-1} (1 - c\alpha)^{2i} \\ &\le \frac{K^2}{1 - (1 - c\alpha)^2}.\end{aligned}$$

Since

$$1 - (1 - c\alpha)^2 = c\alpha(2 - c\alpha) \ge c\alpha,$$

where $0 < c\alpha < 1$, it follows that

$$\sum_{i=0}^{k-1} \|\bar{H}^i\|_\infty^2 \le \frac{K^2}{c\alpha}. \tag{5}$$

Similarly, for every sequence $H_{i+1}, \dots, H_{k-1} \in \mathrm{co}(\mathcal{H}_\alpha)$,

$$\begin{aligned}&\sum_{i=0}^{k-1} \sup_{H_{i+1},\dots,H_{k-1} \in \mathrm{co}(\mathcal{H}_\alpha)} \|H_{k-1} \cdots H_{i+1}\|_\infty \\ &\quad \le K \sum_{i=0}^{k-1} (1 - c\alpha)^{k-1-i} \\ &\quad = K \sum_{j=0}^{k-1} (1 - c\alpha)^j \\ &\quad \le \frac{K}{1 - (1 - c\alpha)} \\ &\quad = \frac{K}{c\alpha}.\end{aligned}$$

Hence

$$\sum_{i=0}^{k-1} \sup_{H_{i+1},\dots,H_{k-1}\in \mathrm{co}(\mathcal{H}_\alpha)} \|H_{k-1}\cdots H_{i+1}\|_\infty \le \frac{K}{c\alpha}, \tag{6}$$

We now apply the reference-filter decomposition. Fix any reference matrix

$$\bar{H} \in \mathrm{co}(\mathcal{H}_\alpha)$$

and consider the reference system driven by the same martingale noise:

$$\hat{x}_{k+1} = \bar{H}\hat{x}_k + \alpha w_k, \qquad \hat{x}_0 = x_0.$$

Decompose

$$\hat{x}_k = z_k + y_k,$$

where

$$z_{k+1} = \bar{H} z_k, \qquad z_0 = x_0,$$

and

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

Since $\bar{H} \in \mathrm{co}(\mathcal{H}_\alpha)$, (4) gives

$$\begin{aligned}\|z_k\|_\infty &= \|\bar{H}^k x_0\|_\infty \\ &\le K(1-c\alpha)^k \|x_0\|_\infty \\ &\le K\beta_\alpha^k \|x_0\|_\infty.\end{aligned}$$

We next bound the noise response y_k. Define

$$Y_k := \mathbb{E}[y_k y_k^\top].$$

Because y_k is $\mathcal{F}_k$-measurable and

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

the martingale cross terms vanish, and therefore

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

Unrolling this recursion gives

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

By the noise bound,

$$\begin{aligned}\lambda_{\max}\big(\mathbb{E}[w_j w_j^\top]\big) &\le \mathbb{E}[\|w_j\|_2^2] \\ &\le W_{\max}.\end{aligned}$$

Therefore,

$$\begin{aligned}\lambda_{\max}(Y_k) &\le \alpha^2 W_{\max} \sum_{i=0}^{k-1} \|\bar{H}^i\|_2^2 \\ &\le \alpha^2 W_{\max} 2|\mathcal{S}||\mathcal{A}| \sum_{i=0}^{k-1} \|\bar{H}^i\|_\infty^2 \\ &\le \alpha^2 W_{\max} 2|\mathcal{S}||\mathcal{A}| \frac{K^2}{c\alpha} \\ &= \frac{\alpha W_{\max} 2|\mathcal{S}||\mathcal{A}| K^2}{c},\end{aligned}$$

where we used (5). Since $Y_k \succeq 0$,

$$\mathrm{tr}(Y_k) \le 2|\mathcal{S}||\mathcal{A}|\lambda_{\max}(Y_k) \le \frac{\alpha W_{\max}(2|\mathcal{S}||\mathcal{A}|)^2 K^2}{c}.$$

Consequently,

$$\mathbb{E}[\|y_k\|_\infty] \le \mathbb{E}[\|y_k\|_2] \le \sqrt{\mathbb{E}[\|y_k\|_2^2]} = \sqrt{\mathrm{tr}(Y_k)} \le 2|\mathcal{S}||\mathcal{A}|K\sqrt{\frac{\alpha W_{\max}}{c}}.$$

We now compare the true recursion with the reference filter. Define

$$r_k := x_k - \hat{x}_k.$$

Since the two recursions are driven by the same martingale noise, the stochastic terms cancel pathwise:

$$\begin{aligned} r_{k+1} &= x_{k+1} - \hat{x}_{k+1} \\ &= H_k x_k + \alpha w_k - \bar{H}\hat{x}_k - \alpha w_k \\ &= H_k r_k + (H_k - \bar{H})\hat{x}_k. \end{aligned}$$

Using $\hat{x}_k = z_k + y_k$, decompose

$$r_k = u_k + v_k,$$

where

$$u_{k+1} = H_k u_k + (H_k - \bar{H}) z_k, \qquad u_0 = 0,$$

and

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

For any deterministic policies π, π',

$$\|DP(\Pi^\pi - \Pi^{\pi'})\|_\infty \le 2d_{\max}.$$

Therefore, for any two extreme modes $H_\theta, H_{\theta'} \in \mathcal{H}_\alpha$,

$$\begin{aligned} \|H_\theta - H_{\theta'}\|_\infty &\le \max\{(4+2)\alpha\gamma p_1 d_{\max}, (2+4)\alpha\gamma p_2 d_{\max}\} \\ &\le 6\alpha\gamma p_{\max} d_{\max}. \end{aligned}$$

By convexity, this bound extends to arbitrary $H, \bar{H} \in \mathrm{co}(\mathcal{H}_\alpha)$, yielding

$$\|H - \bar{H}\|_\infty \le 6\gamma p_{\max} d_{\max} \alpha.$$

We first bound u_k. Unrolling its recursion yields

$$u_k = \sum_{i=0}^{k-1} H_{k-1} \cdots H_{i+1}(H_i - \bar{H}) z_i.$$

Therefore,

$$\begin{aligned} \|u_k\|_\infty &\le \sum_{i=0}^{k-1} K(1-c\alpha)^{k-1-i} 6\gamma p_{\max} d_{\max} \alpha K (1-c\alpha)^i \|x_0\|_\infty \\ &= 6\gamma p_{\max} d_{\max} \alpha K^2 k (1-c\alpha)^{k-1} \|x_0\|_\infty \\ &\le 6\gamma p_{\max} d_{\max} \alpha K^2 k \beta_\alpha^{k-1} \|x_0\|_\infty. \end{aligned}$$

We next bound v_k. Unrolling gives

$$v_k = \sum_{i=0}^{k-1} H_{k-1} \cdots H_{i+1}(H_i - \bar{H}) y_i.$$

Hence,

$$\begin{aligned} \mathbb{E}[\|v_k\|_\infty] &\le 6\gamma p_{\max} d_{\max} \alpha \sum_{i=0}^{k-1} \|H_{k-1} \cdots H_{i+1}\|_\infty \, \mathbb{E}[\|y_i\|_\infty] \\ &\le 6\gamma p_{\max} d_{\max} \alpha \frac{K}{c\alpha} 2|\mathcal{S}||\mathcal{A}|K\sqrt{\frac{\alpha W_{\max}}{c}} \\ &= \frac{6\gamma p_{\max} d_{\max} K}{c} 2|\mathcal{S}||\mathcal{A}|K\sqrt{\frac{\alpha W_{\max}}{c}}, \end{aligned}$$

where (6) was used.

Finally,

$$x_k = \hat{x}_k + r_k = z_k + y_k + u_k + v_k.$$

Combining the preceding estimates gives

$$\begin{aligned}\mathbb{E}[\|x_k\|_\infty] &\le \|z_k\|_\infty + \mathbb{E}[\|y_k\|_\infty] + \|u_k\|_\infty + \mathbb{E}[\|v_k\|_\infty] \\ &\le K\beta_\alpha^k \|x_0\|_\infty \\ &\quad + 6\gamma p_{\max} d_{\max} \alpha K^2 k \beta_\alpha^{k-1} \|x_0\|_\infty \\ &\quad + 2|\mathcal{S}||\mathcal{A}|K\left(1 + \frac{6\gamma p_{\max} d_{\max} K}{c}\right)\sqrt{\frac{\alpha W_{\max}}{c}}.\end{aligned}$$

Because

$$\|x_k\|_\infty = \max\left\{\|Q_k^1 - Q^*\|_\infty, \|Q_k^2 - Q^*\|_\infty\right\},$$

the claimed finite-time bound follows.

Since $0 < \beta_\alpha < 1$,

$$\beta_\alpha^k \to 0, \qquad k\beta_\alpha^{k-1} \to 0.$$

Taking $\limsup_{k\to\infty}$ therefore gives

$$\limsup_{k\to\infty} \mathbb{E}[\|x_k\|_\infty] \le 2|\mathcal{S}||\mathcal{A}|K\left(1 + \frac{6\gamma p_{\max} d_{\max} K}{c}\right)\sqrt{\frac{\alpha W_{\max}}{c}}.$$

All constants in the right-hand side are independent of α, so the noise floor is $O(\sqrt{\alpha})$.

Finally,

$$\begin{aligned}\|\bar{Q}_k - Q^*\|_\infty &= \left\|\frac{(Q_k^1 - Q^*) + (Q_k^2 - Q^*)}{2}\right\|_\infty \\ &\le \max\left\{\|Q_k^1 - Q^*\|_\infty, \|Q_k^2 - Q^*\|_\infty\right\}.\end{aligned}$$

Thus the same bound holds for the averaged estimate. □

B.8 Proof of Proposition 2

Proposition 2 (restated). Consider the AT-DQ augmented family $\widetilde{\mathcal{H}}_{\alpha,\beta}^{\mathrm{AT}}$. Suppose that the tracking step-size β is proportional to the Bellman-update step-size α, and define $\tau := \frac{\beta}{\alpha}$. Assume that $\tau > \gamma\left(\frac{d_{\max}}{d_{\min}} + \frac{(1+\gamma)d_{\max}^2}{(1-\gamma)d_{\min}^2}\right)$. Choose $\omega \in \left(\frac{\gamma d_{\max}}{d_{\min}(1-\gamma)}, \frac{d_{\min}\tau - \gamma d_{\max}}{(1+\gamma)d_{\max}}\right)$, and define $c := \min\left\{d_{\min}(1-\gamma) - \frac{\gamma d_{\max}}{\omega},\ d_{\min}\tau - \gamma d_{\max} - \omega(1+\gamma)d_{\max}\right\}$. Then $c > 0$. Let $\bar{\alpha} := \min\left\{\frac{1}{d_{\max}}, \frac{1}{\tau d_{\max}}, \frac{1}{c}\right\}$. Then, for every $0 < \alpha \le \bar{\alpha}$ with $\beta = \tau\alpha$, the transformed weighted block sup norm $\left\|\begin{bmatrix}\tilde{e}_k^1 \\ \tilde{e}_k^2\end{bmatrix}\right\|_\omega := \max\left\{\|\tilde{e}_k^1\|_\infty, \omega\|\tilde{e}_k^2\|_\infty\right\}$ satisfies $\left\|\begin{bmatrix}\tilde{e}_{k+1}^1 \\ \tilde{e}_{k+1}^2\end{bmatrix}\right\|_\omega \le (1 - \alpha c)\left\|\begin{bmatrix}\tilde{e}_k^1 \\ \tilde{e}_k^2\end{bmatrix}\right\|_\omega$ for every extreme mode of the AT-DQ switching family. The same contraction bound holds for all matrices in $\mathrm{co}(\widetilde{\mathcal{H}}_{\alpha,\beta}^{\mathrm{AT}})$. Consequently, $\rho_{\mathrm{jsr}}\left(\mathrm{co}(\widetilde{\mathcal{H}}_{\alpha,\beta}^{\mathrm{AT}})\right) = \rho_{\mathrm{jsr}}\left(\widetilde{\mathcal{H}}_{\alpha,\beta}^{\mathrm{AT}}\right) \le 1 - \alpha c < 1$.

We work in the transformed coordinates $\tilde{x}_k = [(\tilde{e}_k^1)^\top, (\tilde{e}_k^2)^\top]^\top$. Taking the infinity norm of the transformed deterministic recursion yields the following one-step comparison inequalities for every admissible mode:

$$\begin{aligned}\|\tilde{e}_{k+1}^1\|_\infty &\le \left(1 - \alpha d_{\min}(1-\gamma)\right)\|\tilde{e}_k^1\|_\infty + \alpha\gamma d_{\max}\|\tilde{e}_k^2\|_\infty, \\ \|\tilde{e}_{k+1}^2\|_\infty &\le \alpha(1+\gamma)d_{\max}\|\tilde{e}_k^1\|_\infty + \left(1 - \beta d_{\min} + \alpha\gamma d_{\max}\right)\|\tilde{e}_k^2\|_\infty.\end{aligned}$$

Define the weighted block max norm

$$\|\tilde{x}_k\|_\omega := \max\left\{\|\tilde{e}_k^1\|_\infty, \omega\|\tilde{e}_k^2\|_\infty\right\}.$$

Using the one-step comparison inequalities and $\beta = \tau\alpha$, we obtain

$$\|\tilde{e}_{k+1}^1\|_\infty \le \left(1 - \alpha d_{\min}(1-\gamma) + \frac{\alpha\gamma d_{\max}}{\omega}\right)\|\tilde{x}_k\|_\omega,$$

$$\omega\|\tilde{e}_{k+1}^2\|_\infty \le (1 - \alpha d_{\min}\tau + \alpha\gamma d_{\max} + \alpha\omega(1+\gamma)d_{\max})\,\|\tilde{x}_k\|_\omega.$$

To make both contraction margins strictly positive, we require

$$d_{\min}(1-\gamma) - \frac{\gamma d_{\max}}{\omega} > 0, \qquad d_{\min}\tau - \gamma d_{\max} - \omega(1+\gamma)d_{\max} > 0,$$

or equivalently,

$$\frac{\gamma d_{\max}}{d_{\min}(1-\gamma)} < \omega < \frac{d_{\min}\tau - \gamma d_{\max}}{(1+\gamma)d_{\max}}.$$

To meet this requirement, the above interval must be nonempty, which is equivalent to

$$\tau > \gamma\left(\frac{d_{\max}}{d_{\min}} + \frac{(1+\gamma)d_{\max}^2}{(1-\gamma)d_{\min}^2}\right)$$

implies

$$\frac{\gamma d_{\max}}{d_{\min}(1-\gamma)} < \frac{d_{\min}\tau - \gamma d_{\max}}{(1+\gamma)d_{\max}},$$

so this interval is nonempty. Hence, choose any

$$\omega \in \left(\frac{\gamma d_{\max}}{d_{\min}(1-\gamma)}, \frac{d_{\min}\tau - \gamma d_{\max}}{(1+\gamma)d_{\max}}\right).$$

Now define

$$c := \min\left\{d_{\min}(1-\gamma) - \frac{\gamma d_{\max}}{\omega}, d_{\min}\tau - \gamma d_{\max} - \omega(1+\gamma)d_{\max}\right\}.$$

Therefore, every AT-DQ extreme mode satisfies

$$\|\tilde{x}_{k+1}\|_\omega \le (1-\alpha c)\|\tilde{x}_k\|_\omega.$$

By convexity, the same contraction bound holds for every matrix in $\mathrm{co}(\widetilde{\mathcal{H}}_{\alpha,\beta}^{\mathrm{AT}})$. Indeed, if $\widetilde{H} = \sum_j \lambda_j \widetilde{H}_j$, where $\widetilde{H}_j \in \widetilde{\mathcal{H}}_{\alpha,\beta}^{\mathrm{AT}}$, $\lambda_j \ge 0$, and $\sum_j \lambda_j = 1$, then, for every $\tilde{x} \in \mathbb{R}^{2|\mathcal{S}||\mathcal{A}|}$,

$$\|\widetilde{H}\tilde{x}\|_\omega = \left\|\sum_j \lambda_j \widetilde{H}_j\tilde{x}\right\|_\omega \le \sum_j \lambda_j\|\widetilde{H}_j\tilde{x}\|_\omega \le (1-\alpha c)\|\tilde{x}\|_\omega.$$

Consequently, for every product

$$\widetilde{H}_{j-1}\cdots\widetilde{H}_0, \qquad \widetilde{H}_i \in \mathrm{co}(\widetilde{\mathcal{H}}_{\alpha,\beta}^{\mathrm{AT}}),$$

we have

$$\|\widetilde{H}_{j-1}\cdots\widetilde{H}_0\tilde{x}\|_\omega \le (1-\alpha c)^j\|\tilde{x}\|_\omega.$$

Thus, in the induced ω-norm,

$$\|\widetilde{H}_{j-1}\cdots\widetilde{H}_0\|_\omega \le (1-\alpha c)^j.$$

Taking the supremum over all products of length j, taking the j-th root, and then taking the limit superior gives

$$\rho_{\mathrm{jsr}}(\mathrm{co}(\widetilde{\mathcal{H}}_{\alpha,\beta}^{\mathrm{AT}})) \le 1 - \alpha c.$$

Since the joint spectral radius is invariant under convexification of a finite matrix family,

$$\rho_{\mathrm{jsr}}(\mathrm{co}(\widetilde{\mathcal{H}}_{\alpha,\beta}^{\mathrm{AT}})) = \rho_{\mathrm{jsr}}(\widetilde{\mathcal{H}}_{\alpha,\beta}^{\mathrm{AT}}).$$

Therefore, for every

$$0 < \alpha \le \bar{\alpha} \qquad \text{with} \qquad \beta = \tau\alpha,$$

where

$$\bar{\alpha} := \min\left\{\frac{1}{d_{\max}}, \frac{1}{\tau d_{\max}}, \frac{1}{c}\right\},$$

we have

$$\rho_{\mathrm{jsr}}\left(\widetilde{\mathcal{H}}_{\alpha,\beta}^{\mathrm{AT}}\right) \le 1 - \alpha c < 1.$$

This completes the proof.

C FINITE-TIME ANALYSIS OF AT-DQ

While the main text establishes the JSR stability of AT-DQ, this section provides its finite-time analysis. To this end, we first prove the boundedness of the AT-DQ iterates.

Lemma 10. *Suppose that the one-step rewards satisfy $|r_{k+1}| \le R_{\max}$. The discount factor is $\gamma \in (0,1)$, and the step-sizes satisfy $0 < \alpha, \beta \le 1$ with $\beta = \tau\alpha$ for a given tracking ratio $\tau > 0$ for all $k \ge 0$. Define the uniform invariant bound B and the noise variance bound $W_{\max}$ as:*

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

Then, along every sample path, the iterates of AT-DQ remain uniformly bounded for all $k \ge 0$:

$$\|Q_k^1\|_\infty \le B, \qquad \|Q_k^2\|_\infty \le B.$$

Furthermore, the conditional second moment of the augmented martingale difference noise w_k is uniformly bounded by:

$$\mathbb{E}\left[\|w_k\|_2^2 \mid \mathcal{F}_k\right] \le W_{\max}.$$

Proof. We first prove the pathwise boundedness by mathematical induction. For the base case $k = 0$, we have $\|Q_0^1\|_\infty \le B$ and $\|Q_0^2\|_\infty \le B$ by the definition of B. Assume that for some $k \ge 0$, the iterates satisfy $\|Q_k^1\|_\infty \le B$ and $\|Q_k^2\|_\infty \le B$.

Let $i_k = (s_k, a_k)$ be the sampled state-action pair at iteration k. The Bellman target for the primary table is defined as $Y_k := r_{k+1} + \gamma Q_k^2\left(s'_k, \pi_{Q_k^1}(s'_k)\right)$. By the inductive hypothesis and the definition $B \ge \frac{R_{\max}}{1-\gamma}$, we can bound the target as follows:

$$\begin{aligned}|Y_k| &\le |r_{k+1}| + \gamma \left\|Q_k^2\right\|_\infty \\ &\le R_{\max} + \gamma B \\ &\le (1-\gamma)B + \gamma B = B.\end{aligned}$$

The primary table is updated only at the sampled coordinate i_k. Since $\alpha \in (0,1]$, the update is a convex combination of two terms bounded by B:

$$\begin{aligned}\left|Q_{k+1}^1(i_k)\right| &= \left|(1-\alpha)Q_k^1(i_k) + \alpha Y_k\right| \\ &\le (1-\alpha)\left\|Q_k^1\right\|_\infty + \alpha|Y_k| \\ &\le (1-\alpha)B + \alpha B = B.\end{aligned}$$

Similarly, the tracking table Q^2 is updated at i_k. Since $\beta \in (0,1]$, this is also a convex combination:

$$\begin{aligned}\left|Q_{k+1}^2(i_k)\right| &= \left|(1-\beta)Q_k^2(i_k) + \beta Q_k^1(i_k)\right| \\ &\le (1-\beta)\left\|Q_k^2\right\|_\infty + \beta\left\|Q_k^1\right\|_\infty \\ &\le (1-\beta)B + \beta B = B.\end{aligned}$$

Because all other coordinates remain unchanged (i.e., $Q_{k+1}^1(s,a) = Q_k^1(s,a)$ and $Q_{k+1}^2(s,a) = Q_k^2(s,a)$ for $(s,a) \ne i_k$), it holds that $\|Q_{k+1}^1\|_\infty \le B$ and $\|Q_{k+1}^2\|_\infty \le B$. This completes the induction, proving uniform boundedness along every sample path.

Next, we bound the conditional second moment of the augmented martingale difference noise w_k^{AT}. Let Z_k denote the uncentered stochastic update vector at iteration k, before scaling by the common step-size α. Factoring out α (and using $\beta = \tau\alpha$), the non-zero components of Z_k at the sampled coordinate i_k are:

$$\begin{aligned}Z_k^1(i_k) &= Y_k - Q_k^1(i_k), \\ Z_k^2(i_k) &= \tau\left(Q_k^1(i_k) - Q_k^2(i_k)\right).\end{aligned}$$

Using the established uniform bound B, we can bound the magnitude of each component:

$$\begin{aligned}|Z_k^1(i_k)| &\le |Y_k| + |Q_k^1(i_k)| \le B + B = 2B, \\ |Z_k^2(i_k)| &\le \tau\left(|Q_k^1(i_k)| + |Q_k^2(i_k)|\right) \le \tau(B+B) = 2\tau B.\end{aligned}$$

Thus, the squared Euclidean norm of Z_k is bounded almost surely by:

$$\begin{aligned}\|Z_k\|_2^2 &= |Z_k^1(i_k)|^2 + |Z_k^2(i_k)|^2 \\ &\le (2B)^2 + (2\tau B)^2 \\ &= 4B^2(1+\tau^2).\end{aligned}$$

The augmented martingale difference noise is defined as the conditional centering of this update, i.e., $w_k = Z_k - \mathbb{E}[Z_k \mid \mathcal{F}_k]$. Since conditional centering minimizes the conditional second moment, we obtain:

$$\mathbb{E}\left[\|w_k\|_2^2 \mid \mathcal{F}_k\right] \le \mathbb{E}\left[\|Z_k\|_2^2 \mid \mathcal{F}_k\right] \le 4B^2(1+\tau^2) = W_{max}.$$

This completes the proof. □

Lemma 11. *Under the assumptions of Proposition 2, for every $0 < \alpha \le \bar{\alpha}$ with $\beta = \tau\alpha$,*

$$\rho_{\text{jsr}}\left(\mathcal{H}_{\alpha,\beta}^{\text{AT}}\right) \le 1 - c\alpha < 1.$$

Moreover,

$$\sup_{H_0,\dots,H_{\ell-1}\in\text{co}(\mathcal{H}_{\alpha,\beta}^{\text{AT}})} \|H_{\ell-1}\cdots H_0\|_\infty \le \kappa_\omega(1-c\alpha)^\ell, \qquad \forall \ell \ge 0, \tag{7}$$

where κ_ω is the norm-equivalence constant associated with the transformed weighted block norm.

Proof. Let us define the coordinate transformation $x = T\tilde{x}$, where the original augmented error x and the tracking error $\tilde{x}$ are given by

$$x = \begin{bmatrix} Q^1 - Q^* \\ Q^2 - Q^* \end{bmatrix}, \qquad \tilde{x} = \begin{bmatrix} Q^1 - Q^* \\ Q^2 - Q^1 \end{bmatrix}.$$

The transformation matrix T and its inverse T^{-1} can be explicitly written as

$$T = \begin{bmatrix} I & 0 \\ I & I \end{bmatrix}, \qquad T^{-1} = \begin{bmatrix} I & 0 \\ -I & I \end{bmatrix}.$$

Under this transformation, any matrix $H \in \text{co}(\mathcal{H}_{\alpha,\beta}^{\text{AT}})$ in the original coordinates is represented as $\widetilde{H} = T^{-1}HT$. Based on the weighted norm contraction established previously, for every $\tilde{x} \in \mathbb{R}^{2|\mathcal{S}||\mathcal{A}|}$, any product of these transformed matrices satisfies

$$\left\|\widetilde{H}_{\ell-1}\cdots\widetilde{H}_0\tilde{x}\right\|_\omega \le (1-c\alpha)^\ell\|\tilde{x}\|_\omega.$$

To translate this contraction back to the original ∞-norm coordinates, we evaluate the norm bounds for T and T^{-1}. By the structure of the transformation, the following inequalities hold:

$$\|T\tilde{x}\|_\infty \le \left(1 + \frac{1}{\omega}\right)\|\tilde{x}\|_\omega, \qquad \|T^{-1}x\|_\omega \le \max\{1, 2\omega\}\|x\|_\infty.$$

To see why these bounds hold, write

$$\tilde{x} = \begin{bmatrix} \tilde{e}^1 \\ \tilde{e}^2 \end{bmatrix}.$$

Since

$$T\tilde{x} = \begin{bmatrix} \tilde{e}^1 \\ \tilde{e}^1 + \tilde{e}^2 \end{bmatrix},$$

we have

$$\begin{aligned}\|T\tilde{x}\|_\infty &= \max\left\{\|\tilde{e}^1\|_\infty, \|\tilde{e}^1 + \tilde{e}^2\|_\infty\right\} \\ &\le \max\left\{\|\tilde{e}^1\|_\infty, \|\tilde{e}^1\|_\infty + \|\tilde{e}^2\|_\infty\right\}.\end{aligned}$$

By the definition

$$\|\tilde{x}\|_\omega = \max\left\{\|\tilde{e}^1\|_\infty, \omega\|\tilde{e}^2\|_\infty\right\},$$

we have

$$\|\tilde{e}^1\|_\infty \le \|\tilde{x}\|_\omega, \qquad \|\tilde{e}^2\|_\infty \le \frac{1}{\omega}\|\tilde{x}\|_\omega,$$

and therefore

$$\|T\tilde{x}\|_\infty \le \left(1+\frac{1}{\omega}\right)\|\tilde{x}\|_\omega.$$

Similarly, for

$$x=\begin{bmatrix}e^1\\e^2\end{bmatrix}, \qquad T^{-1}x=\begin{bmatrix}e^1\\e^2-e^1\end{bmatrix},$$

we obtain

$$\begin{aligned}\|T^{-1}x\|_\omega &= \max\left\{\|e^1\|_\infty,\, \omega\|e^2-e^1\|_\infty\right\}\\ &\le \max\left\{\|e^1\|_\infty,\, \omega\left(\|e^2\|_\infty+\|e^1\|_\infty\right)\right\}.\end{aligned}$$

Since

$$\|e^1\|_\infty,\, \|e^2\|_\infty \le \|x\|_\infty,$$

it follows that

$$\|T^{-1}x\|_\omega \le \max\{1,2\omega\}\|x\|_\infty.$$

Since the matrix product in the original coordinates can be factorized as $H_{\ell-1}\cdots H_0 = T\widetilde{H}_{\ell-1}\cdots\widetilde{H}_0T^{-1}$, we can bound its norm for any vector x as follows:

$$\begin{aligned}\|H_{\ell-1}\cdots H_0x\|_\infty &= \left\|T\widetilde{H}_{\ell-1}\cdots\widetilde{H}_0T^{-1}x\right\|_\infty\\ &\le \left(1+\frac{1}{\omega}\right)\left\|\widetilde{H}_{\ell-1}\cdots\widetilde{H}_0T^{-1}x\right\|_\omega\\ &\le \left(1+\frac{1}{\omega}\right)(1-\alpha c)^\ell\|T^{-1}x\|_\omega\\ &\le \left(1+\frac{1}{\omega}\right)\max\{1,2\omega\}(1-\alpha c)^\ell\|x\|_\infty.\end{aligned}$$

By explicitly defining the coordinate transformation constant κ_ω as

$$\kappa_\omega := \left(1+\frac{1}{\omega}\right)\max\{1,2\omega\},$$

we obtain the final uniform product bound:

$$\sup_{H_0,\dots,H_{\ell-1}\in\mathrm{co}(\mathcal{H}^{\mathrm{AT}}_{\alpha,\beta})}\|H_{\ell-1}\cdots H_0\|_\infty \le \kappa_\omega(1-c\alpha)^\ell.$$

Furthermore, taking the ℓ-th root of this product bound and letting $\ell\to\infty$, the constant $\kappa_\omega^{1/\ell}$ vanishes. By the definition of the JSR and its invariance under similarity transformations, this directly implies $\rho_{\mathrm{jsr}}(\mathcal{H}^{\mathrm{AT}}_{\alpha,\beta}) = \rho_{\mathrm{jsr}}(\widetilde{\mathcal{H}}^{\mathrm{AT}}_{\alpha,\beta}) \le 1-c\alpha < 1$. This completes the proof. □

Lemma 12. *Under the assumptions of Lemma 11, define*

$$\beta_\alpha := 1-\frac{c}{2}\alpha.$$

Then

$$\rho_{\mathrm{jsr}}\left(\mathcal{H}^{\mathrm{AT}}_{\alpha,\beta}\right) < \beta_\alpha < 1,$$

and the corresponding product-growth constant satisfies

$$K_{\beta_\alpha} \le \kappa_\omega, \qquad 0<\alpha\le\bar{\alpha}.$$

Proof. By Lemma 11,

$$\rho_{\mathrm{jsr}}\left(\mathcal{H}^{\mathrm{AT}}_{\alpha,\beta}\right) \le 1-c\alpha < 1-\frac{c}{2}\alpha = \beta_\alpha.$$

Moreover, since $\alpha\le\bar{\alpha}\le 1/c$, we have $0<\beta_\alpha<1$.

By the definition of the product-growth constant and (7),

$$K_{\beta_\alpha} = \sup_{\ell \geq 0} \beta_\alpha^{-\ell} \sup_{H_0,\dots,H_{\ell-1} \in \mathrm{co}(\mathcal{H}^{\mathrm{AT}}_{\alpha,\beta})} \|H_{\ell-1} \cdots H_0\|_\infty$$
$$\leq \kappa_\omega \sup_{\ell \geq 0} \left(\frac{1-c\alpha}{\beta_\alpha}\right)^\ell = \kappa_\omega,$$

because

$$0 \leq \frac{1-c\alpha}{1-\frac{c}{2}\alpha} < 1.$$

This completes the proof. □

Theorem 3. *Let Assumption 1 hold. Consider AT-DQ with $\beta = \tau\alpha$, where*

$$\tau > \gamma \left(\frac{d_{\max}}{d_{\min}} + \frac{(1+\gamma)d_{\max}^2}{(1-\gamma)d_{\min}^2}\right).$$

Let ω, c, $\bar{\alpha}$ be as in Proposition 2 and κ_ω be as in Lemma 11, and let W_{max} be as in Lemma 10. For every

$$0 < \alpha \leq \bar{\alpha}, \qquad \beta = \tau\alpha \leq 1,$$

define

$$\beta_\alpha := 1 - \frac{c}{2}\alpha.$$

Then, for all $k \geq 0$,

$$\mathbb{E}\left[\max\left\{\|Q_k^1 - Q^*\|_\infty, \|Q_k^2 - Q^*\|_\infty\right\}\right]$$
$$\leq \kappa_\omega \beta_\alpha^k \max\left\{\|Q_0^1 - Q^*\|_\infty, \|Q_0^2 - Q^*\|_\infty\right\}$$
$$+ 6\gamma d_{\max}\alpha\kappa_\omega^2 k \beta_\alpha^{k-1} \max\left\{\|Q_0^1 - Q^*\|_\infty, \|Q_0^2 - Q^*\|_\infty\right\}$$
$$+ 2|\mathcal{S}||\mathcal{A}|\kappa_\omega \left(1 + \frac{6\gamma d_{\max}\kappa_\omega}{c}\right)\sqrt{\frac{W_{max}\alpha}{c}}.$$

Here $k\beta_\alpha^{k-1}$ is interpreted as zero when $k = 0$. Consequently,

$$\limsup_{k\to\infty} \mathbb{E}\left[\max\left\{\|Q_k^1 - Q^*\|_\infty, \|Q_k^2 - Q^*\|_\infty\right\}\right]$$
$$\leq 2|\mathcal{S}||\mathcal{A}|\kappa_\omega \left(1 + \frac{6\gamma d_{\max}\kappa_\omega}{c}\right)\sqrt{\frac{W_{max}\alpha}{c}}.$$

In particular, the constant-step-size noise floor is $O(\sqrt{\alpha})$.

Proof. Recall that

$$x_k = \begin{bmatrix} Q_k^1 - Q^* \\ Q_k^2 - Q^* \end{bmatrix}.$$

Fix $0 < \alpha \leq \bar{\alpha}$ with $\beta = \tau\alpha \leq 1$. By Lemma 11,

$$\sup_{H_0,\dots,H_{\ell-1} \in \mathrm{co}(\mathcal{H}^{\mathrm{AT}}_{\alpha,\beta})} \|H_{\ell-1} \cdots H_0\|_\infty \leq \kappa_\omega (1-c\alpha)^\ell. \tag{8}$$

For every fixed $\bar{H} \in \mathrm{co}(\mathcal{H}^{\mathrm{AT}}_{\alpha,\beta})$, (8) gives

$$\sum_{i=0}^{k-1} \|\bar{H}^i\|_\infty^2 \leq \kappa_\omega^2 \sum_{i=0}^{k-1} (1-c\alpha)^{2i}$$
$$\leq \frac{\kappa_\omega^2}{c\alpha},$$

where we used

$$1 - (1-c\alpha)^2 \geq c\alpha.$$

Similarly,

$$\sum_{i=0}^{k-1} \sup_{H_{i+1},\dots,H_{k-1}\in \mathrm{co}(\mathcal{H}^{\mathrm{AT}}_{\alpha,\beta})} \|H_{k-1}\cdots H_{i+1}\|_\infty \le \frac{\kappa_\omega}{c\alpha}. \tag{9}$$

Fix $\bar{H} \in \mathrm{co}(\mathcal{H}^{\mathrm{AT}}_{\alpha,\beta})$ and consider

$$\hat{x}_{k+1} = \bar{H}\hat{x}_k + \alpha w_k, \qquad \hat{x}_0 = x_0.$$

Write

$$\hat{x}_k = z_k + y_k,$$

where

$$z_{k+1} = \bar{H} z_k, \qquad z_0 = x_0,$$

and

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

By (8),

$$\|z_k\|_\infty \le \kappa_\omega \beta_\alpha^k \|x_0\|_\infty.$$

By Lemma 10,

$$\mathbb{E}\big[\|w_k\|_2^2 \mid \mathcal{F}_k\big] \le W_{max}.$$

Using the martingale-difference property and the preceding square-summability bound gives

$$\mathbb{E}[\|y_k\|_\infty] \le 2|\mathcal{S}||\mathcal{A}|\kappa_\omega \sqrt{\frac{W_{max}\alpha}{c}}.$$

For any $H, \bar{H} \in \mathrm{co}(\mathcal{H}^{\mathrm{AT}}_{\alpha,\beta})$, the block structure of the AT-DQ modes yields

$$\|H - \bar{H}\|_\infty \le 6\gamma d_{\max}\alpha.$$

Indeed,

$$\|DP(\Pi^\pi - \Pi^{\pi'})\|_\infty \le 2d_{\max},$$

and only the two policy-dependent blocks in the first block row differ between switching modes.

Define

$$r_k := x_k - \hat{x}_k.$$

Since both recursions are driven by the same noise,

$$r_{k+1} = H_k r_k + (H_k - \bar{H})\hat{x}_k.$$

Using $\hat{x}_k = z_k + y_k$, write $r_k = u_k + v_k$, where

$$u_{k+1} = H_k u_k + (H_k - \bar{H}) z_k, \qquad u_0 = 0,$$

and

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

Unrolling the first recursion and using (8) gives

$$\|u_k\|_\infty \le 6\gamma d_{\max}\alpha\kappa_\omega^2 k \beta_\alpha^{k-1}\|x_0\|_\infty.$$

Likewise, by (9),

$$\mathbb{E}[\|v_k\|_\infty] \le \frac{6\gamma d_{\max}\kappa_\omega}{c}\, 2|\mathcal{S}||\mathcal{A}|\kappa_\omega \sqrt{\frac{W_{max}\alpha}{c}}.$$

Finally,

$$x_k = z_k + y_k + u_k + v_k.$$

Combining the preceding estimates yields

$$\begin{aligned}\mathbb{E}[\|x_k\|_\infty] \le{}& \kappa_\omega \beta_\alpha^k \|x_0\|_\infty \\ &+ 6\gamma d_{\max}\alpha\kappa_\omega^2 k\beta_\alpha^{k-1}\|x_0\|_\infty \\ &+ 2|\mathcal{S}||\mathcal{A}|\kappa_\omega \left(1 + \frac{6\gamma d_{\max}\kappa_\omega}{c}\right)\sqrt{\frac{W_{max}\alpha}{c}}.\end{aligned}$$

Since

$$\|x_k\|_\infty = \max\left\{\|Q_k^1 - Q^*\|_\infty, \|Q_k^2 - Q^*\|_\infty\right\},$$

the finite-time bound follows.

Because $0 < \beta_\alpha < 1$,

$$\beta_\alpha^k \to 0, \qquad k\beta_\alpha^{k-1} \to 0.$$

Taking the limit superior proves the asymptotic bound. Since W_{max}, τ, c, and κ_ω are independent of α, the noise floor is $O(\sqrt{\alpha})$.

Finally,

$$\begin{aligned}\|\bar{Q}_k - Q^*\|_\infty &= \left\| \frac{(Q_k^1 - Q^*) + (Q_k^2 - Q^*)}{2} \right\|_\infty \\ &\leq \max\left\{\|Q_k^1 - Q^*\|_\infty, \|Q_k^2 - Q^*\|_\infty\right\}.\end{aligned}$$

Thus the same bound holds for the averaged estimate. □

D ASYMMETRIC GREEDY-SELECTION TRACKING VARIANT

This section presents a closely related greedy-selection tracking variant, which we call Asymmetric Greedy-Selection Tracking double Q-learning (AGT-DQ). In AT-DQ, the primary table selects the greedy action and the tracking table evaluates it. In contrast, AGT-DQ lets the tracking table select the greedy action, while the primary table performs the Bellman evaluation.

At time k, the sampled coordinate (s_k, a_k) is updated by

$$Q^1_{k+1}(s_k,a_k) = Q^1_k(s_k,a_k) + \alpha\left(r_{k+1} + \gamma Q^1_k\big(s'_k, \pi_{Q^2_k}(s'_k)\big) - Q^1_k(s_k,a_k)\right),$$
$$Q^2_{k+1}(s_k,a_k) = Q^2_k(s_k,a_k) + \beta\left(Q^1_k(s_k,a_k) - Q^2_k(s_k,a_k)\right),$$

with all other coordinates unchanged. Here $\alpha > 0$ is the Bellman-update step-size and $\beta > 0$ is the tracking step-size.

Let

$$e^1_k := Q^1_k - Q^*, \qquad e^2_k := Q^2_k - Q^*, \qquad x_k := \begin{bmatrix} e^1_k \\ e^2_k \end{bmatrix}.$$

Write

$$\pi_k := \pi_{Q^2_k}.$$

By Lemma 2, applied with $X = Q^2_k$ and $Y = Q^1_k$, there exists an $\mathcal{F}_k$-measurable stochastic policy μ_k such that

$$\Pi^{\pi_k} Q^1_k - V^* = (\Pi^{\mu_k} - \Pi^{\pi_k}) e^2_k + \Pi^{\pi_k} e^1_k.$$

Therefore the conditional-mean recursion has the form

$$x_{k+1} = H_k x_k + \alpha w_k, \qquad \mathbb{E}[w_k \mid \mathcal{F}_k] = 0,$$

where

$$H_k = \begin{bmatrix} I - \alpha D + \alpha\gamma DP\Pi^{\pi_k} & \alpha\gamma DP(\Pi^{\mu_k} - \Pi^{\pi_k}) \\ \beta D & I - \beta D \end{bmatrix}.$$

Consequently, the deterministic extreme AGT-DQ family is

$$\mathcal{H}^{\mathrm{AGT}}_{\alpha,\beta} := \left\{ H^{\mathrm{AGT}}_{\pi,\nu}(\alpha,\beta) : \pi, \nu \in \Theta \right\},$$

where

$$H^{\mathrm{AGT}}_{\pi,\nu}(\alpha,\beta) := \begin{bmatrix} I - \alpha D + \alpha\gamma DP\Pi^{\pi} & \alpha\gamma DP(\Pi^{\nu} - \Pi^{\pi}) \\ \beta D & I - \beta D \end{bmatrix}.$$

Every H_k belongs to $\mathrm{co}(\mathcal{H}^{\mathrm{AGT}}_{\alpha,\beta})$. The signed policy-difference block $\Pi^\nu - \Pi^\pi$ again prevents a direct nonnegative row-sum contraction argument in the original (e^1, e^2) coordinates. However, the tracking structure yields a simple certificate after the change of variables

$$\tilde{e}^1_k := Q^1_k - Q^*, \qquad \tilde{e}^2_k := Q^2_k - Q^1_k, \qquad \tilde{x}_k := \begin{bmatrix} \tilde{e}^1_k \\ \tilde{e}^2_k \end{bmatrix}.$$

For each extreme mode, the transformed deterministic recursion is

$$\tilde{e}^1_{k+1} = (I - \alpha D + \alpha\gamma DP\Pi^\nu)\tilde{e}^1_k + \alpha\gamma DP(\Pi^\nu - \Pi^\pi)\tilde{e}^2_k,$$

and

$$\tilde{e}^2_{k+1} = \alpha D(I - \gamma P\Pi^\nu)\tilde{e}^1_k + (I - \beta D - \alpha\gamma DP(\Pi^\nu - \Pi^\pi))\,\tilde{e}^2_k.$$

The corresponding deterministic extreme family is $\tilde{\mathcal{H}}^{\mathrm{AGT}}_{\alpha,\beta} := \{\tilde{H}^{\mathrm{AGT}}_{\pi,\nu}(\alpha,\beta) : \pi, \nu \in \Theta\}$, where

$$\tilde{H}^{\mathrm{AGT}}_{\pi,\nu}(\alpha,\beta) := \begin{bmatrix} I - \alpha D + \alpha\gamma DP\Pi^{\nu} & \alpha\gamma DP(\Pi^{\nu} - \Pi^{\pi}) \\ \alpha D(I - \gamma P\Pi^{\nu}) & I - \beta D - \alpha\gamma DP(\Pi^{\nu} - \Pi^{\pi}) \end{bmatrix}.$$

The following proposition shows that the greedy-selection tracking mechanism also induces a weighted block-norm contraction. Hence AGT-DQ admits an analogous JSR stability.

Proposition 3. *Consider the AGT-DQ augmented family $\widetilde{\mathcal{H}}^{\mathrm{AGT}}_{\alpha,\beta}$. Suppose that the tracking step-size β is proportional to the Bellman-update step-size α, and define $\tau := \frac{\beta}{\alpha}$. Assume that $\tau > 2\gamma\left(\frac{d_{\max}}{d_{\min}} + \frac{(1+\gamma)d_{\max}^2}{(1-\gamma)d_{\min}^2}\right)$. Choose $\omega \in \left(\frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)}, \frac{d_{\min}\tau - 2\gamma d_{\max}}{(1+\gamma)d_{\max}}\right)$, and define $c := \min\left\{d_{\min}(1-\gamma) - \frac{2\gamma d_{\max}}{\omega},\ d_{\min}\tau - 2\gamma d_{\max} - \omega(1+\gamma)d_{\max}\right\}$. Then $c > 0$. Let $\bar{\alpha} := \min\left\{\frac{1}{d_{\max}}, \frac{1}{\tau d_{\max}}, \frac{1}{c}\right\}$. Then, for every $0 < \alpha \le \bar{\alpha}$ with $\beta = \tau\alpha$, the transformed weighted block sup norm $\left\|\begin{bmatrix}\tilde{e}_k^1\\ \tilde{e}_k^2\end{bmatrix}\right\|_\omega := \max\left\{\|\tilde{e}_k^1\|_\infty, \omega\|\tilde{e}_k^2\|_\infty\right\}$ satisfies $\left\|\begin{bmatrix}\tilde{e}_{k+1}^1\\ \tilde{e}_{k+1}^2\end{bmatrix}\right\|_\omega \le (1-\alpha c)\left\|\begin{bmatrix}\tilde{e}_k^1\\ \tilde{e}_k^2\end{bmatrix}\right\|_\omega$ for every extreme mode of the AGT-DQ switching family. The same contraction bound holds for all matrices in $\mathrm{co}(\widetilde{\mathcal{H}}^{\mathrm{AGT}}_{\alpha,\beta})$. Consequently,*

$$\rho_{\mathrm{jsr}}\Big(\mathrm{co}(\widetilde{\mathcal{H}}^{\mathrm{AGT}}_{\alpha,\beta})\Big) = \rho_{\mathrm{jsr}}\Big(\widetilde{\mathcal{H}}^{\mathrm{AGT}}_{\alpha,\beta})\Big) \le 1 - \alpha c < 1.$$

Proof. We work in the transformed coordinates $\tilde{x}_k = [(\tilde{e}_k^1)^\top, (\tilde{e}_k^2)^\top]^\top$. Taking the infinity norm of the transformed deterministic recursion yields the following one-step comparison inequalities for every admissible mode under $0 < \alpha, \beta \le 1/d_{\max}$:

$$\begin{aligned}\|\tilde{e}_{k+1}^1\|_\infty &\le (1 - \alpha d_{\min}(1-\gamma))\,\|\tilde{e}_k^1\|_\infty + 2\alpha\gamma d_{\max}\|\tilde{e}_k^2\|_\infty,\\ \|\tilde{e}_{k+1}^2\|_\infty &\le \alpha(1+\gamma)d_{\max}\|\tilde{e}_k^1\|_\infty + (1 - \beta d_{\min} + 2\alpha\gamma d_{\max})\,\|\tilde{e}_k^2\|_\infty.\end{aligned}$$

Since the tracking ratio $\tau = \beta/\alpha$ satisfies

$$\tau > 2\gamma\left(\frac{d_{\max}}{d_{\min}} + \frac{(1+\gamma)d_{\max}^2}{(1-\gamma)d_{\min}^2}\right),$$

there exists a weight $\omega > 0$ satisfying

$$\frac{2\gamma d_{\max}}{d_{\min}(1-\gamma)} < \omega < \frac{d_{\min}\tau - 2\gamma d_{\max}}{(1+\gamma)d_{\max}}.$$

Define the weighted block max norm

$$\|\tilde{x}_k\|_\omega := \max\left\{\|\tilde{e}_k^1\|_\infty, \omega\|\tilde{e}_k^2\|_\infty\right\}.$$

Using the one-step comparison inequalities and $\beta = \tau\alpha$, we obtain

$$\begin{aligned}\|\tilde{e}_{k+1}^1\|_\infty &\le \left(1 - \alpha d_{\min}(1-\gamma) + \frac{2\alpha\gamma d_{\max}}{\omega}\right)\|\tilde{x}_k\|_\omega,\\ \omega\|\tilde{e}_{k+1}^2\|_\infty &\le (1 - \alpha d_{\min}\tau + 2\alpha\gamma d_{\max} + \alpha\omega(1+\gamma)d_{\max})\,\|\tilde{x}_k\|_\omega.\end{aligned}$$

Define

$$c := \min\left\{d_{\min}(1-\gamma) - \frac{2\gamma d_{\max}}{\omega},\ d_{\min}\tau - 2\gamma d_{\max} - \omega(1+\gamma)d_{\max}\right\}.$$

By the choice of ω, we have $c > 0$. Therefore every AGT-DQ transformed homogeneous mode satisfies

$$\|\tilde{x}_{k+1}\|_\omega \le (1 - \alpha c)\|\tilde{x}_k\|_\omega.$$

By convexity, the same contraction bound holds for every matrix in $\mathrm{co}(\widetilde{\mathcal{H}}^{\mathrm{AGT}}_{\alpha,\beta})$. Indeed, if $\widetilde{H} = \sum_j \lambda_j \widetilde{H}_j$, where $\widetilde{H}_j \in \widetilde{\mathcal{H}}^{\mathrm{AGT}}_{\alpha,\beta}$, $\lambda_j \ge 0$, and $\sum_j \lambda_j = 1$, then, for every $\tilde{x} \in \mathbb{R}^{2|\mathcal{S}||\mathcal{A}|}$,

$$\|\widetilde{H}\tilde{x}\|_\omega = \left\|\sum_j \lambda_j \widetilde{H}_j\tilde{x}\right\|_\omega \le \sum_j \lambda_j\|\widetilde{H}_j\tilde{x}\|_\omega \le (1-\alpha c)\|\tilde{x}\|_\omega.$$

Consequently, for every product

$$\widetilde{H}_{j-1}\cdots\widetilde{H}_0, \qquad \widetilde{H}_i \in \mathrm{co}(\widetilde{\mathcal{H}}^{\mathrm{AGT}}_{\alpha,\beta}),$$

we have

$$\|\widetilde{H}_{j-1}\cdots\widetilde{H}_0\tilde{x}\|_\omega \le (1-\alpha c)^j\|\tilde{x}\|_\omega.$$

Taking the supremum over all products of length j, taking the j-th root, and then taking the limit superior gives

$$\rho_{\mathrm{jsr}}\left(\mathrm{co}(\widetilde{\mathcal{H}}^{\mathrm{AGT}}_{\alpha,\beta})\right) \le 1-\alpha c.$$

Since the JSR is invariant under convexification of a finite matrix family,

$$\rho_{\mathrm{jsr}}\left(\mathrm{co}(\widetilde{\mathcal{H}}^{\mathrm{AGT}}_{\alpha,\beta})\right) = \rho_{\mathrm{jsr}}\left(\widetilde{\mathcal{H}}^{\mathrm{AGT}}_{\alpha,\beta}\right).$$

Therefore, for every

$$0<\alpha\le\bar{\alpha} \qquad \text{with} \qquad \beta=\tau\alpha,$$

where

$$\bar{\alpha}_{\mathrm{AGT}} := \min\left\{\frac{1}{d_{\max}}, \frac{1}{\tau d_{\max}}, \frac{1}{c}\right\},$$

we have

$$\rho_{\mathrm{jsr}}\left(\widetilde{\mathcal{H}}^{\mathrm{AGT}}_{\alpha,\beta}\right) \le 1-\alpha c < 1.$$

This completes the proof.

□

D.1 Finite-time analysis of AGT-DQ

Similar to the analysis of AT-DQ, we now present the finite-time analysis for the AGT-DQ variant. We begin by establishing the uniform boundedness of its iterates.

Lemma 13. *Suppose that the one-step rewards satisfy $|r_{k+1}| \le R_{\max}$. The discount factor is $\gamma\in(0,1)$, and the step-sizes satisfy $0<\alpha,\beta\le 1$ with $\beta=\tau\alpha$ for a given tracking ratio $\tau>0$ for all $k\ge 0$. Define the uniform invariant bound B and the noise variance bound $W_{\max}$ as:*

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

Then, along every sample path, the iterates of AGT-DQ remain uniformly bounded for all $k\ge 0$:

$$\|Q_k^1\|_\infty \le B, \qquad \|Q_k^2\|_\infty \le B.$$

Furthermore, the conditional second moment of the augmented martingale difference noise w_k is uniformly bounded by:

$$\mathbb{E}\left[\|w_k\|_2^2 \mid \mathcal{F}_k\right] \le W_{\max}.$$

Proof. We prove the pathwise boundedness by induction. The claim holds at $k=0$ by the definition of B. Suppose that

$$\|Q_k^1\|_\infty \le B, \qquad \|Q_k^2\|_\infty \le B.$$

Let $i_k=(s_k,a_k)$, and define

$$Y_k := r_{k+1} + \gamma Q_k^1\big(s_k', \pi_{Q_k^2}(s_k')\big).$$

Then

$$\begin{aligned}|Y_k| &\le R_{\max} + \gamma\|Q_k^1\|_\infty \\ &\le R_{\max} + \gamma B \le B.\end{aligned}$$

Since $0<\alpha\le 1$,

$$\begin{aligned}|Q_{k+1}^1(i_k)| &= \left|(1-\alpha)Q_k^1(i_k) + \alpha Y_k\right| \\ &\le (1-\alpha)B + \alpha B = B.\end{aligned}$$

Likewise, since $0<\beta\le 1$,

$$\begin{aligned}|Q_{k+1}^2(i_k)| &= \left|(1-\beta)Q_k^2(i_k) + \beta Q_k^1(i_k)\right| \\ &\le (1-\beta)B + \beta B = B.\end{aligned}$$

All other coordinates remain unchanged, which completes the induction.

Let Z_k denote the uncentered stochastic update before scaling by α. Its nonzero components at i_k are

$$Z_k^1(i_k) = Y_k - Q_k^1(i_k), \qquad Z_k^2(i_k) = \tau\big(Q_k^1(i_k) - Q_k^2(i_k)\big).$$

Therefore,

$$|Z_k^1(i_k)| \le 2B, \qquad |Z_k^2(i_k)| \le 2\tau B,$$

and hence

$$\|Z_k\|_2^2 \le 4B^2(1+\tau^2).$$

Since

$$w_k = Z_k - \mathbb{E}[Z_k \mid \mathcal{F}_k],$$

conditional centering gives

$$\begin{aligned}\mathbb{E}\big[\|w_k\|_2^2 \mid \mathcal{F}_k\big] &\le \mathbb{E}\big[\|Z_k\|_2^2 \mid \mathcal{F}_k\big] \\ &\le 4B^2(1+\tau^2) = W_{max}.\end{aligned}$$

□

Lemma 14. *Under the assumptions of Proposition 3, for every $0 < \alpha \le \bar{\alpha}$ with $\beta = \tau\alpha$,*

$$\rho_{\mathrm{jsr}}\big(\mathcal{H}_{\alpha,\beta}^{\mathrm{AGT}}\big) \le 1 - c\alpha < 1.$$

Moreover,

$$\sup_{H_0,\dots,H_{\ell-1}\in \mathrm{co}(\mathcal{H}_{\alpha,\beta}^{\mathrm{AGT}})} \|H_{\ell-1}\cdots H_0\|_\infty \le \kappa_\omega(1-c\alpha)^\ell, \qquad \forall \ell \ge 0, \tag{10}$$

where κ_ω is the norm-equivalence constant associated with the transformed weighted block norm.

Proof. Let us define the coordinate transformation $x = T\tilde{x}$, where the original augmented error x and the tracking error $\tilde{x}$ are given by

$$x = \begin{bmatrix} Q^1 - Q^* \\ Q^2 - Q^* \end{bmatrix}, \qquad \tilde{x} = \begin{bmatrix} Q^1 - Q^* \\ Q^2 - Q^1 \end{bmatrix}.$$

The transformation matrix T and its inverse T^{-1} can be explicitly written as

$$T = \begin{bmatrix} I & 0 \\ I & I \end{bmatrix}, \qquad T^{-1} = \begin{bmatrix} I & 0 \\ -I & I \end{bmatrix}.$$

Under this transformation, any matrix $H \in \mathrm{co}(\mathcal{H}_{\alpha,\beta}^{\mathrm{AGT}})$ in the original coordinates is represented as $\widetilde{H} = T^{-1}HT$. Based on the weighted norm contraction established previously, any product of these transformed matrices satisfies

$$\left\|\widetilde{H}_{\ell-1}\cdots\widetilde{H}_0\tilde{x}\right\|_\omega \le (1-c\alpha)^\ell\|\tilde{x}\|_\omega.$$

To translate this contraction back to the original ∞-norm coordinates, we evaluate the norm bounds for T and T^{-1}. By the structure of the transformation, the following inequalities hold:

$$\|T\tilde{x}\|_\infty \le \left(1+\frac{1}{\omega}\right)\|\tilde{x}\|_\omega, \qquad \|T^{-1}x\|_\omega \le \max\{1,2\omega\}\|x\|_\infty.$$

Since the matrix product in the original coordinates can be factorized as $H_{\ell-1}\cdots H_0 = T\widetilde{H}_{\ell-1}\cdots\widetilde{H}_0T^{-1}$, we can bound its norm for any vector x as follows:

$$\begin{aligned}\|H_{\ell-1}\cdots H_0x\|_\infty &= \left\|T\widetilde{H}_{\ell-1}\cdots\widetilde{H}_0T^{-1}x\right\|_\infty \\ &\le \left(1+\frac{1}{\omega}\right)\left\|\widetilde{H}_{\ell-1}\cdots\widetilde{H}_0T^{-1}x\right\|_\omega \\ &\le \left(1+\frac{1}{\omega}\right)(1-\alpha c)^\ell\|T^{-1}x\|_\omega \\ &\le \left(1+\frac{1}{\omega}\right)\max\{1,2\omega\}(1-\alpha c)^\ell\|x\|_\infty.\end{aligned}$$

By explicitly defining the coordinate transformation constant κ_ω as

$$\kappa_\omega := \left(1 + \frac{1}{\omega}\right) \max\{1, 2\omega\},$$

we obtain the final uniform product bound:

$$\sup_{H_0,\dots,H_{\ell-1} \in \mathrm{co}(\mathcal{H}^{\mathrm{AGT}}_{\alpha,\beta})} \|H_{\ell-1} \cdots H_0\|_\infty \le \kappa_\omega (1 - c\alpha)^\ell.$$

This completes the proof. □

Lemma 15. *Under the assumptions of Lemma 14, define*

$$\beta_\alpha := 1 - \frac{c}{2}\alpha.$$

Then

$$\rho_{\mathrm{jsr}}\big(\mathcal{H}^{\mathrm{AGT}}_{\alpha,\beta}\big) < \beta_\alpha < 1,$$

and the corresponding product-growth constant satisfies

$$K_{\beta_\alpha} \le \kappa_\omega, \qquad 0 < \alpha \le \bar{\alpha}.$$

Proof. By Lemma 14,

$$\begin{aligned}\rho_{\mathrm{jsr}}\big(\mathcal{H}^{\mathrm{AGT}}_{\alpha,\beta}\big) &\le 1 - c\alpha \\ &< 1 - \frac{c}{2}\alpha = \beta_\alpha.\end{aligned}$$

Moreover, $\alpha \le \bar{\alpha} \le 1/c$ implies $0 < \beta_\alpha < 1$.

By (10),

$$\begin{aligned}K_{\beta_\alpha} &= \sup_{\ell \ge 0} \beta_\alpha^{-\ell} \sup_{H_0,\dots,H_{\ell-1} \in \mathrm{co}(\mathcal{H}^{\mathrm{AGT}}_{\alpha,\beta})} \|H_{\ell-1} \cdots H_0\|_\infty \\ &\le \kappa_\omega \sup_{\ell \ge 0} \left(\frac{1 - c\alpha}{\beta_\alpha}\right)^\ell \\ &= \kappa_\omega,\end{aligned}$$

because

$$0 \le \frac{1 - c\alpha}{1 - \frac{c}{2}\alpha} < 1.$$

□

Theorem 4. *Let Assumption 1 hold. Consider AGT-DQ with*

$$\beta = \tau\alpha,$$

where

$$\tau > 2\gamma \left(\frac{d_{\max}}{d_{\min}} + \frac{(1+\gamma)d_{\max}^2}{(1-\gamma)d_{\min}^2}\right).$$

Let ω, c, $\bar{\alpha}$, and κ_ω be as in Proposition 3, and let W_{max} be as in Lemma 13. For every

$$0 < \alpha \le \bar{\alpha}, \qquad \beta = \tau\alpha \le 1,$$

define

$$\beta_\alpha := 1 - \frac{c}{2}\alpha.$$

Then, for all $k \ge 0$,

$$\begin{aligned}&\mathbb{E}\big[\max\{\|Q_k^1 - Q^*\|_\infty, \|Q_k^2 - Q^*\|_\infty\}\big] \\ &\quad\le \kappa_\omega \beta_\alpha^k \max\{\|Q_0^1 - Q^*\|_\infty, \|Q_0^2 - Q^*\|_\infty\} \\ &\qquad + 6\gamma d_{\max} \alpha \kappa_\omega^2 k \beta_\alpha^{k-1} \max\{\|Q_0^1 - Q^*\|_\infty, \|Q_0^2 - Q^*\|_\infty\} \\ &\qquad + 2|\mathcal{S}||\mathcal{A}|\kappa_\omega \left(1 + \frac{6\gamma d_{\max}\kappa_\omega}{c}\right) \sqrt{\frac{W_{max}\alpha}{c}}.\end{aligned}$$

Here $k\beta_\alpha^{k-1}$ is interpreted as zero when $k = 0$. Consequently,

$$\limsup_{k\to\infty} \mathbb{E}\big[\max\{\|Q_k^1 - Q^*\|_\infty, \|Q_k^2 - Q^*\|_\infty\}\big] \le 2|\mathcal{S}||\mathcal{A}|\kappa_\omega \left(1 + \frac{6\gamma d_{\max}\kappa_\omega}{c}\right)\sqrt{\frac{W_{max}\alpha}{c}}.$$

Proof. Recall that

$$x_k = \begin{bmatrix} Q_k^1 - Q^* \\ Q_k^2 - Q^* \end{bmatrix}.$$

By Lemma 14,

$$\sup_{H_0,\dots,H_{\ell-1}\in \mathrm{co}(\mathcal{H}_{\alpha,\beta}^{\mathrm{AGT}})} \|H_{\ell-1}\cdots H_0\|_\infty \le \kappa_\omega(1-c\alpha)^\ell. \tag{11}$$

Moreover,

$$K_{\beta_\alpha} \le \kappa_\omega$$

by Lemma 15.

For every fixed $\bar{H} \in \mathrm{co}(\mathcal{H}_{\alpha,\beta}^{\mathrm{AGT}})$,

$$\sum_{i=0}^{k-1} \|\bar{H}^i\|_\infty^2 \le \kappa_\omega^2 \sum_{i=0}^{k-1}(1-c\alpha)^{2i} \le \frac{\kappa_\omega^2}{c\alpha},$$

and

$$\sum_{i=0}^{k-1} \sup_{H_{i+1},\dots,H_{k-1}\in \mathrm{co}(\mathcal{H}_{\alpha,\beta}^{\mathrm{AGT}})} \|H_{k-1}\cdots H_{i+1}\|_\infty \le \frac{\kappa_\omega}{c\alpha}. \tag{12}$$

Fix $\bar{H} \in \mathrm{co}(\mathcal{H}_{\alpha,\beta}^{\mathrm{AGT}})$ and consider

$$\hat{x}_{k+1} = \bar{H}\hat{x}_k + \alpha w_k, \qquad \hat{x}_0 = x_0.$$

Write

$$\hat{x}_k = z_k + y_k,$$

where

$$z_{k+1} = \bar{H} z_k, \qquad z_0 = x_0,$$

and

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

Then

$$\|z_k\|_\infty \le \kappa_\omega \beta_\alpha^k \|x_0\|_\infty.$$

By Lemma 13,

$$\mathbb{E}\big[\|w_k\|_2^2 \mid \mathcal{F}_k\big] \le W_{max}.$$

The martingale-difference property and the square-summability bound therefore give

$$\mathbb{E}[\|y_k\|_\infty] \le 2|\mathcal{S}||\mathcal{A}|\kappa_\omega\sqrt{\frac{W_{max}\alpha}{c}}.$$

For every $H, \bar{H} \in \mathrm{co}(\mathcal{H}_{\alpha,\beta}^{\mathrm{AGT}})$,

$$\|H - \bar{H}\|_\infty \le 6\gamma d_{\max}\alpha.$$

Indeed, the difference between two extreme modes satisfies

$$\left\|\alpha\gamma DP(\Pi^\pi - \Pi^{\pi'})\right\|_\infty \le 2\alpha\gamma d_{\max}$$

in the first diagonal block, and

$$\left\|\alpha\gamma DP\big[(\Pi^\nu - \Pi^\pi) - (\Pi^{\nu'} - \Pi^{\pi'})\big]\right\|_\infty \le 4\alpha\gamma d_{\max}$$

in the upper-right block.

Define

$$r_k := x_k - \hat{x}_k.$$

Because the two recursions are driven by the same noise,

$$r_{k+1} = H_k r_k + (H_k - \bar{H})\hat{x}_k.$$

Using $\hat{x}_k = z_k + y_k$, write $r_k = u_k + v_k$, where

$$u_{k+1} = H_k u_k + (H_k - \bar{H}) z_k, \qquad u_0 = 0,$$

and

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

Unrolling the first recursion gives

$$\|u_k\|_\infty \le 6\gamma d_{\max} \alpha \kappa_\omega^2 k \beta_\alpha^{k-1} \|x_0\|_\infty.$$

Likewise, using (12),

$$\mathbb{E}[\|v_k\|_\infty] \le \frac{6\gamma d_{\max} \kappa_\omega}{c} 2|\mathcal{S}||\mathcal{A}|\kappa_\omega \sqrt{\frac{W_{max}\alpha}{c}}.$$

Finally,

$$x_k = z_k + y_k + u_k + v_k.$$

Combining the preceding estimates proves the finite-time bound.

Since

$$\beta_\alpha^k \to 0, \qquad k\beta_\alpha^{k-1} \to 0,$$

taking the limit superior gives the stated noise-floor bound. The constants are independent of α, so the noise floor is $O(\sqrt{\alpha})$.

Finally,

$$\begin{aligned} \|\bar{Q}_k - Q^*\|_\infty &= \left\| \frac{(Q_k^1 - Q^*) + (Q_k^2 - Q^*)}{2} \right\|_\infty \\ &\le \max\left\{ \|Q_k^1 - Q^*\|_\infty, \|Q_k^2 - Q^*\|_\infty \right\}. \end{aligned}$$

Thus the same bound holds for the averaged estimate.

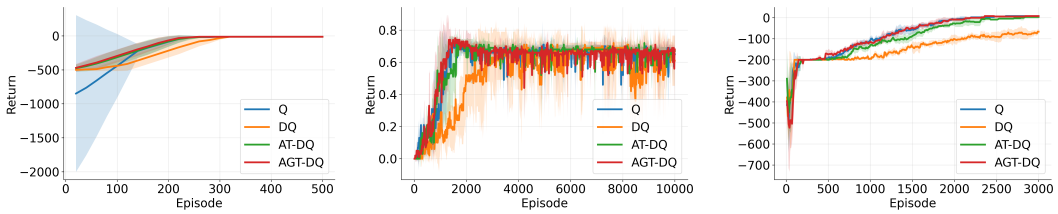
□

D.2 Additional Experiments for AGT-DQ

We conduct additional experiments for AGT-DQ using the same experimental protocol as in the main paper. Specifically, we evaluate AGT-DQ on both the convergence benchmarks and the maximization bias benchmarks. The Bellman-update step-size and the tracking step-size are set to the same values as those used for AT-DQ.

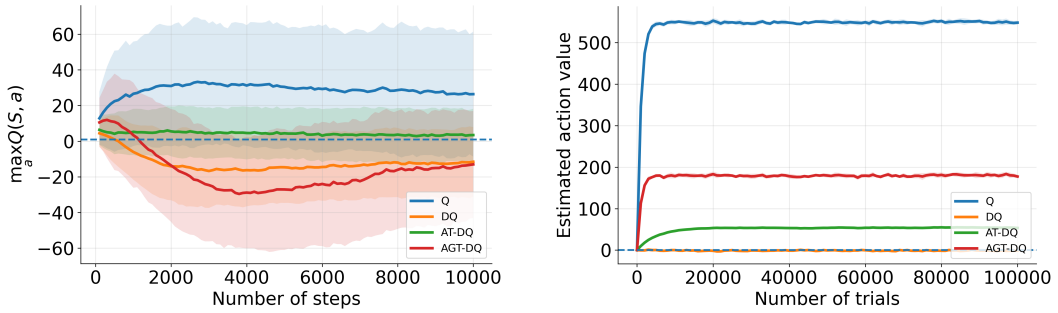
As shown in Figure 2, AGT-DQ exhibits stronger convergence performance than double Q-learning. This result suggests that the greedy-selection tracking mechanism also improves the stability of the learning dynamics under constant step sizes.

However, the maximization bias results in Figure 3 show a different trend. Although AGT-DQ reduces the maximization bias compared with standard Q-learning, it exhibits larger bias than both AT-DQ and double Q-learning. Thus, AGT-DQ is effective from the viewpoint of convergence stabilization, but its bias-reduction effect is weaker than that of AT-DQ and double Q-learning.



(a) CliffWalking (b) FrozenLake (c) Taxi

Figure 2: Convergence performance comparison of Q-learning, double Q-learning, AT-DQ, and AGT-DQ on three benchmark environments.



(a) Start-state value estimates in the GridWorld environment.

(b) Maximization bias in the Roulette environment.

Figure 3: Comparison of maximization bias behavior of Q-learning, double Q-learning, AT-DQ, and AGT-DQ on the GridWorld and Roulette benchmarks. The GridWorld experiment reports the evolution of the start-state value estimate, while the Roulette experiment reports the mean estimated value of betting actions.

E Explicit Mode Calculation for Example 1

In this section, we explicitly construct the mode matrix G_θ used in Example 1 and verify that it has a positive eigenvalue, which implies $\rho_\alpha > 1$ for every $0 < \alpha < 1$.

Under the state-action ordering

$$(1,1),(2,1),(3,1),(1,2),(2,2),(3,2),$$

the sampling matrix is

$$D = \operatorname{diag}\left(0.44,\ 0.0625,\ 0.343,\ 0.0035,\ 0.11,\ 0.041\right),$$

and the transition matrix is

$$P = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

For the mode

$$\theta = (\pi_1, \nu_1, \pi_2, \nu_2),$$

we use

$$\big(\pi_1(1), \pi_1(2), \pi_1(3)\big) = (1,1,2), \qquad \big(\nu_1(1), \nu_1(2), \nu_1(3)\big) = (2,2,1).$$

Hence,

$$\Pi^{\pi_1} = \Pi^{\pi_2} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix},$$

and

$$\Pi^{\nu_1} = \Pi^{\nu_2} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \end{bmatrix}.$$

Using $p_1 = 0.91$, $p_2 = 0.09$, and $\gamma = 0.995$, the mode matrix is

$$G_\theta = \begin{bmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{bmatrix},$$

where

$$G_{11} = -p_1 D + \gamma p_1 DP\left(\Pi^{\nu_1} - \Pi^{\pi_1}\right), \qquad G_{12} = \gamma p_1 DP\Pi^{\pi_1},$$
$$G_{21} = \gamma p_2 DP\Pi^{\pi_2}, \qquad G_{22} = -p_2 D + \gamma p_2 DP\left(\Pi^{\nu_2} - \Pi^{\pi_2}\right).$$

Numerically,

$$G_{11} = \begin{bmatrix} -0.400400 & 0 & 0.398398 & 0 & 0 & -0.398398 \\ -0.056590625 & -0.056875 & 0 & 0.056590625 & 0 & 0 \\ 0 & -0.31056935 & -0.31213 & 0 & 0.31056935 & 0 \\ -0.003169075 & 0 & 0 & -0.000015925 & 0 & 0 \\ 0 & -0.0995995 & 0 & 0 & -0.0005005 & 0 \\ -0.03712345 & 0 & 0 & 0.03712345 & 0 & -0.03731 \end{bmatrix},$$

$$G_{12} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0.398398 \\ 0.056590625 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.31056935 & 0 & 0 & 0 & 0 \\ 0.003169075 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0995995 & 0 & 0 & 0 & 0 \\ 0.03712345 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

$$G_{21} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0.039402 \\ 0.005596875 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.03071565 & 0 & 0 & 0 & 0 \\ 0.000313425 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.0098505 & 0 & 0 & 0 & 0 \\ 0.00367155 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

and

$$G_{22} = \begin{bmatrix} -0.039600 & 0 & 0.039402 & 0 & 0 & -0.039402 \\ -0.005596875 & -0.005625 & 0 & 0.005596875 & 0 & 0 \\ 0 & -0.03071565 & -0.03087 & 0 & 0.03071565 & 0 \\ -0.000313425 & 0 & 0 & -0.000001575 & 0 & 0 \\ 0 & -0.0098505 & 0 & 0 & -0.0000495 & 0 \\ -0.00367155 & 0 & 0 & 0.00367155 & 0 & -0.00369 \end{bmatrix}.$$

The eigenvalues of G_θ are approximately

$$\begin{aligned} \mathrm{eig}(G_\theta) = \{&0.091442330777,\ -2.85742531 \times 10^{-6},\ -8.96536918 \times 10^{-5}, \\ &-0.001544370909,\ -0.003611860136,\ -0.012031493052,\ -0.017397135323,\ -0.034490972970, \\ &-0.064575078582 \pm 0.011660249236\,\mathrm{i}, \\ &-0.390095665053 \pm 0.104337361312\,\mathrm{i}\}. \end{aligned}$$

Let

$$\lambda_+ = 0.091442330777 > 0.$$

Since

$$H_\theta(\alpha) = I_{12} + \alpha G_\theta,$$

$1 + \alpha\lambda_+$ is an eigenvalue of $H_\theta(\alpha)$. Thus, for every $0 < \alpha < 1$,

$$\rho(H_\theta(\alpha)) \geq 1 + \alpha\lambda_+ > 1.$$

Consequently,

$$\rho_\alpha \geq \rho(H_\theta(\alpha)) > 1.$$

F SAMPLE-PATH SIMULATION FOR EXAMPLE 1

This appendix complements Example 1 with a sample-path simulation. The purpose of the experiment is not to claim that every sample path of double Q-learning is unstable. Rather, it demonstrates that the mode-wise instability identified in Example 1 can be accompanied by persistently larger estimation errors in the actual double Q-learning recursion.

For the MDP in Example 1, the two double-Q tables are initialized as

$$Q_0^1 = [2261 \quad 649 \quad 1056 \quad 1929 \quad 1433 \quad 1059]^\top,$$

and

$$Q_0^2 = [-1895 \quad 16878 \quad 17736 \quad 437 \quad 1759 \quad 12146]^\top.$$

For comparison, standard Q-learning is initialized at their average, $Q_0 = \frac{Q_0^1+Q_0^2}{2}$.

At each iteration, a state-action pair is sampled i.i.d. according to the distribution specified in Example 1, and the updated double-Q table is independently selected according to $p_1 = 0.91$ and $p_2 = 0.09$. Standard Q-learning uses the same sampled state-action sequence and ignores the table-selection sequence.

We define the step-wise ℓ_∞ errors as

$$E_k^{\mathrm{DQ}} := \left\| \frac{Q_k^1 + Q_k^2}{2} - Q^* \right\|_\infty, \quad E_k^{\mathrm{Q}} := \|Q_k - Q^*\|_\infty.$$

To reduce short-term fluctuations, the trajectories are summarized using window-mean ℓ_∞ errors. For each step-size α, the number of iterations is chosen as

$$K_\alpha \approx \frac{10^5}{\alpha}.$$

Thus, smaller step sizes are simulated for proportionally more iterations. This ensures that all experiments cover approximately the same normalized time horizon, where

$$t = \alpha k, \qquad t_{\max} = 10^5.$$

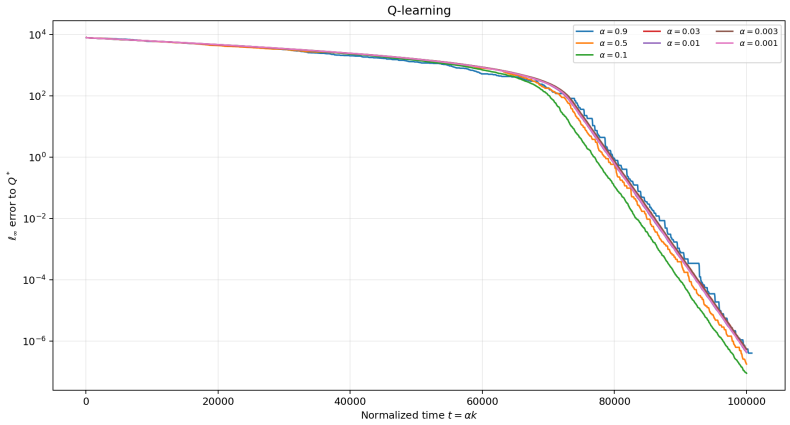
Accordingly, the horizontal axis of the trajectory plots represents the normalized time $t = \alpha k$ rather than the raw iteration index k.

The trajectory is divided into 2000 equal windows, and the error within each window w is averaged as

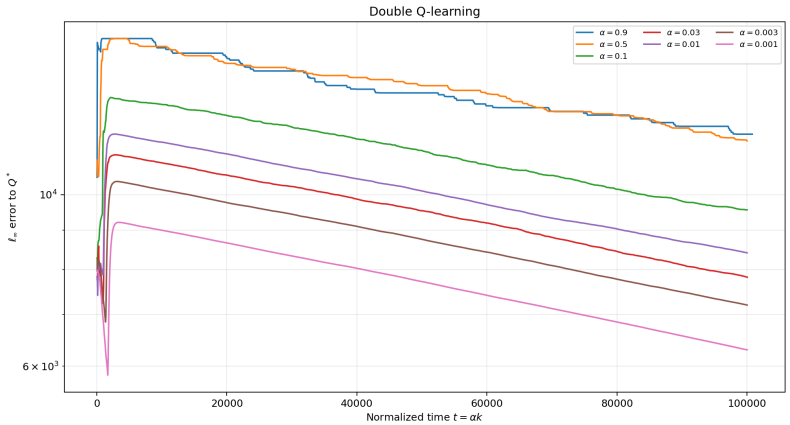
$$\bar{E}_w^{\mathrm{alg}} := \frac{1}{|w|} \sum_{k \in w} E_k^{\mathrm{alg}}, \qquad \mathrm{alg} \in \{\mathrm{Q}, \mathrm{DQ}\}.$$

Representative error trajectories are illustrated in Figure 4. To characterize the long-run behavior, we record the final error of the iterates over the last 25% of the simulation.

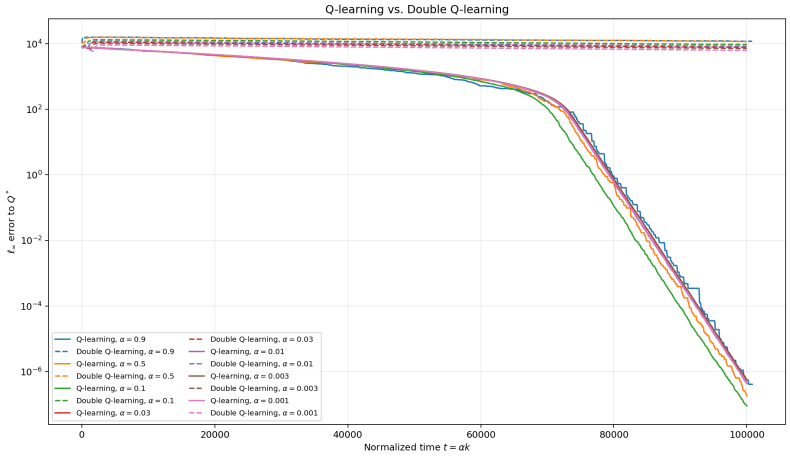
Over the evaluated range of constant step sizes, standard Q-learning approaches Q^*, whereas double Q-learning exhibits substantially larger error and remains far from Q^*. These results provide empirical evidence that the identified mode-wise instability is not merely an artifact of the augmented switching analysis and can manifest in actual double Q-learning trajectories.



(a) Q-learning



(b) Double Q-learning



(c) Comparison of Q-learning and Double Q-learning

Figure 4: Error trajectories of Q-learning and double Q-learning under the same stochastic sampling setting for different constant step sizes. The horizontal axis represents the normalized stochastic-approximation time $t = \alpha k$, and the vertical axis represents the window-mean ℓ_∞ error to Q^*. Standard Q-learning approaches Q^*, whereas double Q-learning exhibits persistently large estimation errors. The bottom panel directly compares the two algorithms.

Table 1: Final ℓ_∞ errors of Q-learning and double Q-learning over five independent random seeds. Each entry reports the mean $\pm$ sample standard deviation across seeds.

α	Q-learning	Double Q-learning
0.9	$(1.77 \pm 3.90) \times 10^{-4}$	3015.93 ± 5037.16
0.5	$(1.07 \pm 2.03) \times 10^{-5}$	5758.76 ± 3996.70
0.1	$(9.46 \pm 9.42) \times 10^{-7}$	6753.95 ± 3942.64
0.03	$(5.85 \pm 3.72) \times 10^{-7}$	7024.18 ± 996.55
0.01	$(3.94 \pm 0.96) \times 10^{-7}$	5503.58 ± 3289.43
0.003	$(4.87 \pm 1.26) \times 10^{-7}$	6823.53 ± 738.94
0.001	$(4.41 \pm 0.51) \times 10^{-7}$	6498.51 ± 695.52

G Experimental Details

G.1 Environment Descriptions

G.1.1 Roulette Environment

The Roulette environment is a single-state MDP designed to expose maximization bias under stochastic rewards. At each decision step, the agent chooses either a termination action or one of many betting actions. The termination action yields zero reward and ends the process, while every betting action returns a random reward with negative expected value and transitions back to the same state.

Although all betting actions are suboptimal in expectation, the large reward variance can induce overestimation when action values are updated through a maximization operator. Consequently, the environment serves as a canonical benchmark for evaluating the maximization bias. Following the original double Q-learning study (Hasselt, 2010), we use this environment to compare the maximization bias characteristics of Q-learning, double Q-learning, AT-DQ, and AGT-DQ.

G.1.2 Gridworld

The Gridworld environment is adapted from the benchmark introduced in the original double Q-learning paper (Hasselt, 2010) to illustrate maximization bias under stochastic rewards. The environment consists of a 3×3 grid with a start state located in the lower-left corner and a goal state in the upper-right corner.

At each non-terminal state, the agent can move in one of four cardinal directions. Attempting to move outside the grid leaves the state unchanged. We employ a reward distribution that yields $+50$ or -52 with equal probability. The goal state provides a reward of $+5$ and terminates the episode.

Although the expected non-terminal reward remains negative, the large reward variance amplifies the overestimation effect caused by the maximization operator. Under $\gamma = 0.99$, the optimal start-state action value is

$$\max_a Q^*(S, a) = 5\gamma^4 - \sum_{k=0}^{3} \gamma^k \approx 0.863.$$

This environment therefore provides a controlled benchmark for evaluating the maximization bias.

G.2 Hyperparameter Settings

G.2.1 Convergence

The convergence experiments were conducted using the hyperparameter settings summarized in Tables 2 and 3.

Table 2: Common hyperparameter settings used in the convergence experiments.

Hyperparameter	**Value**
Discount factor (γ)	0.99
Number of seeds	10
Evaluation interval	100 episodes
Evaluation episodes	20
Initial exploration rate	1.0
Final exploration rate	0.05
Exploration decay horizon	3000 episodes
Q-learning stepsize	0.1
Double Q-learning stepsize	0.1
AT-DQ Bellman-update stepsize (α)	0.1
AT-DQ tracking stepsize (β)	0.8
AGT-DQ Bellman-update stepsize (α)	0.1
AGT-DQ tracking stepsize (β)	0.8

Table 3: Environment-specific settings for the convergence experiments.

Environment	**Training Episodes**	**Max Episode Length**
FrozenLake-v1	10,000	N/A
CliffWalking-v0	500	500
Taxi-v3	3,000	N/A

G.2.2 MAXIMIZATION BIAS

The maximization bias experiments were conducted using the hyperparameter settings summarized in Tables 4 and 5.

Table 4: Hyperparameter settings for the maximization bias experiments.

Hyperparameter	**Value**
Discount factor (γ)	0.99
Bellman-update stepsize (α)	0.1
Tracking stepsize (β) for AT-DQ/AGT-DQ	0.025
Ratio (β/α) for AT-DQ/AGT-DQ	0.25
Roulette trials	100,000
Roulette seeds	10
Roulette logging interval	1,000 trials
Gridworld steps	10,000
Gridworld seeds	1,000
Gridworld logging interval	100 steps

Table 5: Environment-specific settings for the maximization bias experiments.

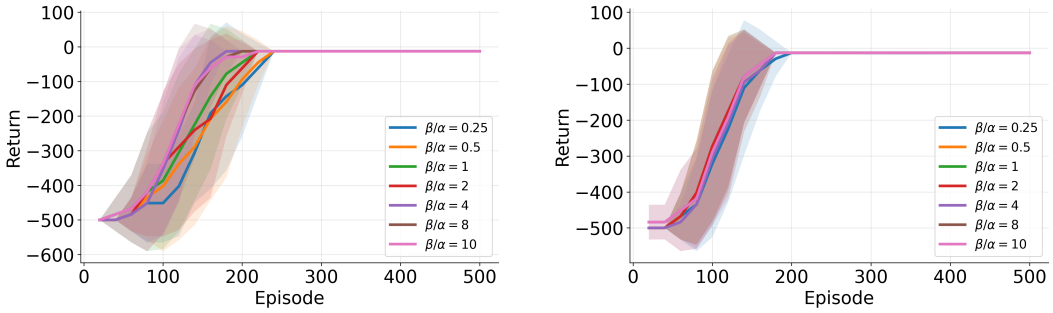
Environment	Reward setting	Metric
Roulette	Win: $+35$, lose: -1, stop: 0	Estimated betting value
Gridworld	Step: $+50/-52$, goal: $+5$	Start-state value bias

H Ablation Study

H.1 Ablation Study on Convergence

In this section, we investigate the effect of the tracking ratio β/α on the convergence behavior of the proposed asymmetric double Q-learning algorithms. We consider the ratios $\beta/\alpha \in \{0.25, 0.5, 1, 2, 4, 8, 10\}$ and evaluate the algorithms on three benchmark environments: CliffWalking, FrozenLake, and Taxi.

Figures 5–7 show the learning curves and final performances obtained under different tracking ratios. As predicted by the tracking dynamics, larger values of β/α improve the convergence speed because the secondary table tracks the primary table more aggressively.



(a) AT-DQ: learning curves for different β/α ratios.

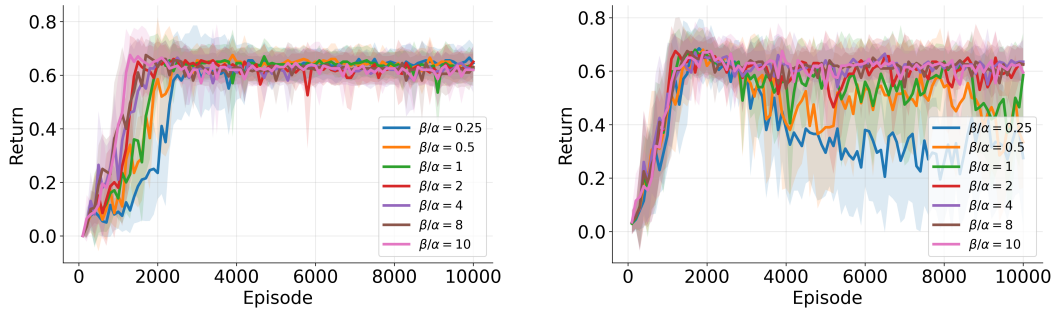
(b) AGT-DQ: learning curves for different β/α ratios.

Figure 5: Ablation study on the effect of the tracking ratio β/α in CliffWalking. The left figure shows the convergence behavior of AT-DQ under different tracking ratios, while the right figure shows the corresponding results for AGT-DQ.

H.2 Ablation Study on Maximization Bias

We investigate the effect of the tracking ratio β/α on maximization bias. As discussed in the main text, the secondary estimator serves as an evaluation table while the primary estimator is responsible for action selection. The tracking ratio therefore controls the degree of separation between the two estimators. To examine this effect, we vary $\beta/\alpha \in \{0.25, 0.5, 1, 2, 4, 8, 10\}$ and evaluate the resulting algorithms on the Roulette and GridWorld maximization bias benchmarks.

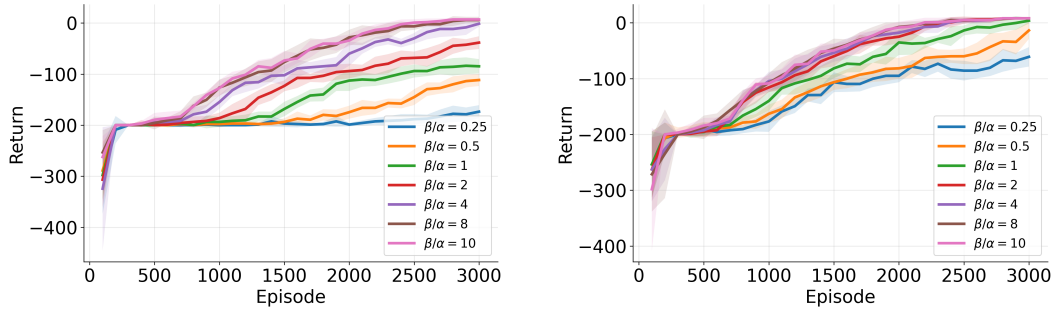
Figures 8 and 9 show that smaller values of β/α generally lead to stronger bias reduction. When β is small, the secondary estimator evolves more slowly, maintaining a larger separation from the primary estimator and thereby increasing the decorrelation between action selection and action evaluation. This effect mitigates the overestimation phenomenon caused by the maximization operator. In contrast, larger values of β/α allow the secondary estimator to track the primary estimator more aggressively, resulting in faster convergence but weaker bias suppression. These results demonstrate a clear trade-off between convergence speed and maximization bias reduction governed by the tracking ratio.



(a) AT-DQ: learning curves for different β/α ratios.

(b) AGT-DQ: learning curves for different β/α ratios.

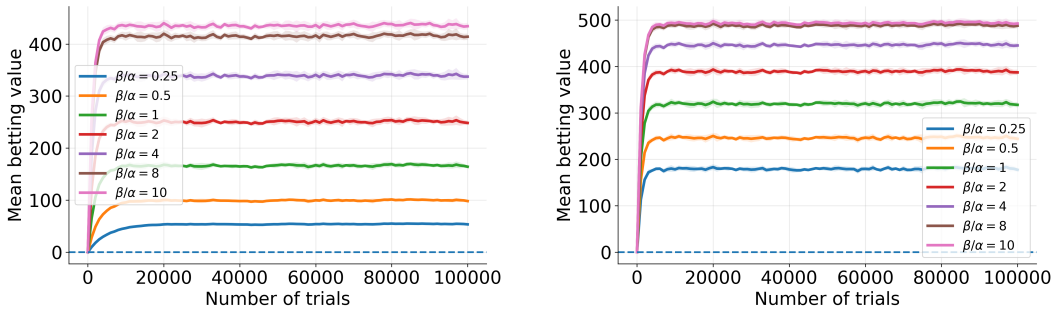
Figure 6: Ablation study on the effect of the tracking ratio β/α in FrozenLake. The left figure shows the convergence behavior of AT-DQ under different tracking ratios, while the right figure shows the corresponding results for AGT-DQ.



(a) AT-DQ: learning curves for different β/α ratios.

(b) AGT-DQ: learning curves for different β/α ratios.

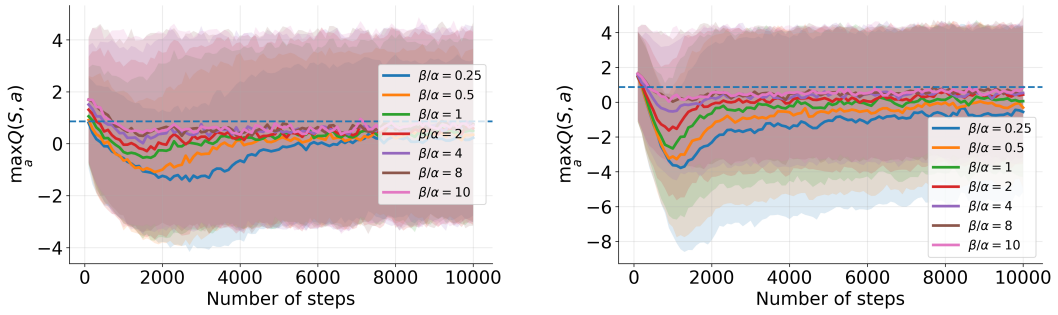
Figure 7: Ablation study on the effect of the tracking ratio β/α in Taxi. The left figure shows the convergence behavior of AT-DQ under different tracking ratios, while the right figure shows the corresponding results for AGT-DQ.



(a) AT-DQ: maximization bias under different β/α ratios.

(b) AGT-DQ: maximization bias under different β/α ratios.

Figure 8: Ablation study on the effect of the tracking ratio β/α for maximization bias reduction in the Roulette environment. The horizontal dashed line indicates the true value of choosing not to bet, which is zero. Smaller values of β/α increase the separation between the selection and evaluation estimators, leading to stronger suppression of maximization bias.



(a) AT-DQ: start-state value estimates under different β/α ratios.

(b) AGT-DQ: start-state value estimates under different β/α ratios.

Figure 9: Ablation study on the effect of the tracking ratio β/α on value estimation in the stochastic GridWorld environment. The figures show the evolution of the start-state value estimate under different tracking ratios.