V-SHARE: PROVABLE CONVERGENCE GUARANTEES FOR ADVANTAGE-BASED ALGORITHMS

Anonymous authors
Paper under double-blind review

ABSTRACT

In reinforcement learning, advantage-based value-learning methods, unlike standard Q-learning, share information across actions by decomposing the action-value function into a state-dependent value component and action-specific advantages. While this mechanism has proven effective in both tabular and deep reinforcement learning, its theoretical understanding remains incomplete. We introduce **V-SHARE**, an advantage-based value-learning algorithm that combines cross-action value sharing with an importance-weighted Bellman-residual correction. To our knowledge, we provide the first complete theoretical characterization of an advantage-based algorithm, including both asymptotic convergence and non-asymptotic finite-sample guarantees. We identify a regime in which V-SHARE mitigates the rare-action bottleneck of asynchronous Q-learning and uncover a contraction–variance tradeoff between stronger deterministic contraction and increased importance-weighting noise. Motivated by this tradeoff, we develop decreasing residual-preconditioning schedules that combine faster early contraction with lower asymptotic noise. We further develop a two-timescale analysis that removes the single-timescale residual-weight threshold while preserving convergence guarantees. Experiments on discrete Markov decision processes (MDPs) and deep reinforcement-learning benchmarks corroborate our results.

1 INTRODUCTION

Reinforcement learning (RL) (Sutton & Barto, 1998) algorithms learn through trial and error: they try actions, receive a reward in return indicating the quality of that action and, over time, try to learn an optimal mapping from the current state to the best action to take, which we call the policy. Q-learning (Watkins & Dayan, 1992) is one of the canonical RL algorithms that learns a Q-function as an intermediate object. The Q-function estimates, for every state-action pair, the long-term return, i.e., the sum of received rewards over a trajectory, obtained by taking that action in that state and behaving optimally thereafter. If the state and action spaces are finite, we can explicitly store estimated action-values for each state-action pair. In the standard asynchronous setting, the learner observes one transition at a time and updates only the estimate for the sampled state-action pair. This simple coordinate-wise structure, together with the contraction of the Bellman optimality operator, underlies the classical convergence theory of Q-learning (Tsitsiklis, 1994; Jaakkola et al., 1993). The same basic principle also proved remarkably successful with function approximation through Deep Q-Networks (DQN) (Mnih et al., 2015), showing that Q-learning can scale to high-dimensional control problems.

With function approximation, however, Q-learning becomes more susceptible to instability, and estimating the full action-value of every action separately can be inefficient when much of the information is shared. These observations motivated the dueling architecture of Wang et al. (2016), which represents the Q-function as the sum of a state-value function that assigns a value to every state defined by the long-term return obtained by behaving optimally after visiting said state, and an action-dependent advantage function learned separately. Intuitively, rather than repeatedly estimating the full value of every action, the learner can first estimate a potentially large value component common to all actions and then learn the comparatively smaller action-dependent advantage components. This value-advantage perspective has subsequently proved useful both empirically and theoretically. Advantage-based architectures improve learning performance in deep reinforcement learning (Wang et al., 2016; Nagarajan et al., 2026), while recent tabular advantage-based algo-

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.

rithms exhibit accelerated learning relative to standard Q-learning (Tang et al., 2023; Daley et al., 2025; Lee, 2026).

The algorithmic idea itself dates back to Advantage Updating (Baird, 1993), and the central mechanism of all these approaches is *value sharing*: updating the state-value component simultaneously changes the represented values of all actions at the state being updated, allowing information obtained from one action to propagate immediately to the others.

Value sharing, however, comes with a theoretical cost. In Q-learning, each state-action value is updated through its own temporal-difference error, which leads to a simple contraction-based analysis. Advantage-based methods couple the action-values within a state through a shared value component, and this coupling breaks that standard argument. Existing analyses therefore rely on stronger structural or sampling assumptions, such as synchronous updates over all state-action pairs, to establish convergence in the tabular setting (Lee, 2026; Tang et al., 2023). This raises a natural question:

Can we retain the acceleration benefits of cross-action value sharing while recovering a tractable convergence theory and finite-sample guarantees under stochastic sampling?

Contributions. We answer this question through V-SHARE, an advantage-based value-learning algorithm that augments cross-action value sharing with a tunable, importance-weighted Bellman-residual correction parametrized by β. This leads to a residual preconditioning view of value sharing, showing when cross-action coupling accelerates learning, when it breaks the usual contraction argument, and how to recover convergence guarantees.

1. We show that the iterates generated by V-SHARE are stable and converge almost surely to the optimal action-value function Q^* under Markovian sampling and suitable parameter choices.
2. We derive finite-time guarantees using constant- and diminishing-stepsize schedules. The resulting rates quantify a contraction-variance tradeoff and yield the first finite-time guarantees of an advantage-based algorithm under Markovian sampling. We show how decreasing-β and decay-to-zero stepsize schedules retain strong early contraction while reducing terminal stochastic error.
3. We also analyze a two-timescale formulation that removes the single-timescale large-β requirement and show convergence guarantees for all $\beta > 0$. The analysis requires extending single-timescale Markovian stochastic approximation to vanishing drift perturbations, yielding results that may be of independent interest.
4. We evaluate V-SHARE and its variants on tabular MDPs, where the experiments expose the predicted effects of value sharing. We additionally extend V-SHARE to deep reinforcement learning and compare it with recent advantage-based methods on MinAtar (Young & Tian, 2019) and Classic Control tasks.

Related work. Since the number of Q-learning variants and analyses is vast, we focus here only on the key theoretical results, major practical milestones, and methods that explicitly learn advantages. The term *advantage-based* is used in several different senses in reinforcement learning. Policy-gradient and actor–critic methods such as PPO (Schulman et al., 2017) use advantage estimates to construct policy updates, often through generalized advantage estimation (Schulman et al., 2015). However, the advantage is used primarily as an estimator of relative action quality to train the policy. Our focus is instead on *advantage-based value learning*: algorithms that explicitly represent action values through a shared state-value component and action-specific advantages, so that updating the value component propagates information across actions. This viewpoint dates back to Advantage Updating (Baird, 1993) and underlies the dueling architecture of Wang et al. (2016).

Recent work has revisited this value–advantage decomposition in both tabular and deep RL (Tang et al., 2023; Daley et al., 2025; Nagarajan et al., 2026). These methods exploit the shared value component to accelerate learning, while Lee (2026) studies the dynamics of tabular dueling Q-learning through spectral arguments. Mustafin et al. (2025) introduces the RB-S algorithm, which uses advantage-preserving transformations to recover an optimal policy. Our approach is complementary: we modify the coupled value-sharing dynamics through residual preconditioning and analyze the resulting algorithm under Markovian sampling, obtaining both asymptotic and finite-time guarantees. Technically, our analysis builds on the ODE method for stochastic approximation (Borkar & Meyn, 2000; Liu et al., 2025) and the generalized Moreau-envelope framework for contractive stochastic approximation (Chen et al., 2020; 2024).

For brevity, we henceforth refer to advantage-based value-learning algorithms as *advantage-based algorithms*.

2 Problem Setup

We first formalize the MDP and sampling model, and then express Q-learning and value-sharing updates that motivate our analysis.

Discounted MDP and sampling model. A **finite discounted Markov decision process (MDP)** is defined by $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action spaces, respectively, $P(s'|s,a)$ is the transition kernel, $r(s,a) \in [0,1]$ is the bounded reward, and $\gamma \in [0,1)$ is the discount factor. The goal is to find a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the expected cumulative discounted reward $\mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r_t]$. For a given MDP, we define the Q-function as $Q(s,a) = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k r_{t+k} \mid s_t = s, a_t = a]$. The **Bellman optimality operator** $\mathcal{T}$ is $(\mathcal{T}Q)(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q(s',a')$. Its unique fixed point is the **optimal action-value function** $Q^\star$.

Synchronous algorithms update all state-action pairs at every iteration, which is computationally expensive and unrealistic in large MDPs. In the asynchronous setting, updates are driven by individual transitions sampled along a trajectory. Our algorithms operate in this regime, learning from a single trajectory generated by a fixed behavior policy π_b. We make a standard assumption about this policy:

Assumption 1 (Ergodic behavior policy)**.** *The behavior policy $\pi_b : \mathcal{S} \to \Delta(\mathcal{A})$ has full support, $\pi_b(a|s) > 0$ for every (s,a), and the state Markov chain induced by π_b is irreducible and aperiodic. Hence it admits a unique stationary distribution $d \in \mathbb{R}^{|\mathcal{S}|}$ satisfying $d(s) > 0$ for every s, and stationary state-action distribution $\mu(s,a) = d(s)\pi_b(a|s)$. Further, let $Y_t := (s_{t-1}, a_{t-1}, s_t)$ denote the Markov chain of observed transitions, with stationary distribution $d_Y(s,a,s') = d(s)\pi_b(a|s)P(s'|s,a)$. For $\delta > 0$, define its mixing time by $t_\delta := \min\{m \geq 0 : \max_y ||\mathbb{P}(Y_m \in \cdot \mid Y_0 = y) - d_Y||_{\mathrm{TV}} \leq \delta\}$, where $||\cdot||_{\mathrm{TV}}$ is the total variation distance. Since Y_t is a finite-state ergodic Markov chain, it is geometrically mixing, and hence $t_\delta = O(\log(1/\delta))$.*

At iteration t, the sampled transition is $(s_t, a_t, r(s_t,a_t), s_{t+1})$, where $a_t \sim \pi_b(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t,a_t)$. The Markovian sample is defined as $Y_{t+1} := (s_t, a_t, s_{t+1})$. For the current iterate Q_t, define the **temporal-difference (TD) error** $\delta_t = r(s_t,a_t) + \gamma \max_{a'} Q_t(s_{t+1},a') - Q_t(s_t,a_t)$. For a stochastic-approximation (SA) recursion $Q_{t+1} = Q_t + \alpha_t \widehat{h}_t(Q_t, Y_{t+1})$, we call $h(Q) := \mathbb{E}_{Y\sim d_Y}[\widehat{h}_t(Q,Y)]$ its **stationary mean operator** and $\dot{Q} = h(Q)$ its associated **limiting ODE** (Borkar, 2022). These dynamics let us compare Q-learning with cross-action value sharing.

Q-learning and value sharing. Asynchronous **Q-learning** with stepsize schedule $\{\alpha_t\}_{t=0}^{\infty}$ updates only the sampled state-action coordinate,

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t \mathbf{1}\{s_t = s, a_t = a\}\delta_t,$$

where $\mathbf{1}$ denotes the indicator function. Let $D_\mu := \mathrm{diag}\{\mu(s,a)\}_{s,a} = \mathrm{diag}\{d(s)\pi_b(a|s)\}_{s,a}$. Its limiting ODE is $\dot{Q} = D_\mu(\mathcal{T}Q - Q)$. Thus asynchronous Q-learning applies a diagonal sampling preconditioner to the **Bellman residual** $\mathcal{T}Q - Q$, preserving the coordinate-wise structure underlying its standard contraction analysis.

We are interested instead in algorithms that share information across actions. To formalize this, consider the **value-advantage decomposition** $Q_t(s,a) = V_t(s) + A_t(s,a)$ and the shared-value update of hard Regularized Dueling Q-learning (RDQ) (Daley et al., 2025),

$$V_{t+1}(s_t) = V_t(s_t) + \alpha_t \delta_t, \quad A_{t+1}(s_t,a) = A_t(s_t,a) + \alpha_t \mathbf{1}\{a_t = a\}\delta_t. \tag{1}$$

Although (V_t, A_t) is not uniquely identifiable, by defining $(LV)(s,a) := V(s)$, the iterate $Q_t = LV_t + A_t$ is well defined.

To understand the effect of value sharing on the Q-dynamics, let $b_s := (\pi_b(a|s))_{a\in\mathcal{A}}$, $\mathbf{e}$ be the all-ones vector in $\mathbb{R}^{|\mathcal{A}|}$, $B_s := \mathbf{e}b_s^\top$, and $B := \mathrm{blkdiag}(B_s : s \in \mathcal{S})$. Also define the lifted state-visitation matrix $D := \mathrm{diag}\{d(s) : (s,a) \in \mathcal{S}\times\mathcal{A}\}$. The Q-iterate then satisfies

$$Q_{t+1}(s,a) - Q_t(s,a) = \alpha_t \mathbf{1}\{s_t = s\}\big(1 + \mathbf{1}\{a_t = a\}\big)\delta_t,$$

with limiting ODE $\dot{Q} = (DB + D_\mu)(\mathcal{T}Q - Q)$.

The distinction from asynchronous Q-learning is the following: The term $D_\mu(\mathcal{T}Q - Q)$ performs the usual action-specific update, whereas $DB(\mathcal{T}Q - Q)$ propagates information across all actions at the visited state. This cross-action coupling is the value-sharing mechanism we study in this paper. Appendix A derives the stationary mean operators of asynchronous Q-learning and hard RDQ and compares the structure of their respective preconditioners and limiting ODEs.

Algorithm 1 V-SHARE

1: Initialize V_0, A_0, state S_0, stepsizes $\{\alpha_t\}_{t=0}^{\infty}$, and choose $\beta > 0$.
2: **for** $t = 0, 1, \ldots$ **do**
3: Sample $a_t \sim \pi_b(\cdot \mid s_t)$.
4: Observe $R_{t+1} = r(s_t, a_t)$ and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$.
5: Compute

$$\delta_t = R_{t+1} + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t).$$

6: Update, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,

$$V_{t+1}(s) = V_t(s) + \mathbf{1}\{s_t = s\}\alpha_t\delta_t, \quad A_{t+1}(s,a) = A_t(s,a) + \alpha_t \frac{\beta\,\mathbf{1}\{s_t = s, a_t = a\}}{\pi_b(a \mid s_t)}\delta_t.$$

7: **end for**

3 Main Results

The value-sharing update equation 1 propagates a sampled TD error across all actions at the visited state, but introduces cross-action coupling that prevents the standard Q-learning contraction argument. We therefore introduce **V-SHARE** in section 3.1 (Algorithm 1), which augments the shared value update with an importance-weighted action-specific Bellman-residual correction. Our analysis casts V-SHARE as a stochastic approximation scheme for a fixed-point iteration. In Section 3.2, we prove almost-sure convergence of V-SHARE and provide finite-time guarantees in Section 3.3. Finally, we investigate convergence properties when we decrease the parameter β of V-SHARE over time in Section 3.4 and under a two-timescale formulation in Section 3.5.

3.1 V-SHARE Algorithm

Let $\beta > 0$ denote the **residual-preconditioning parameter**. Given a sample $(s_t, a_t, r(s_t, a_t), s_{t+1})$ generated by the behaviour policy π_b, V-SHARE performs

$$V_{t+1}(s_t) = V_t(s_t) + \alpha_t\delta_t, \qquad A_{t+1}(s_t, a) = A_t(s_t, a) + \alpha_t \frac{\beta\,\mathbf{1}\{a_t = a\}}{\pi_b(a|s_t)}\delta_t, \tag{2}$$

with $Q_t(s,a) = V_t(s) + A_t(s,a)$ and all unvisited-state coordinates unchanged.

The induced action-value recursion admits the following stochastic approximation (SA) form and limiting ODE,

$$Q_{t+1} = Q_t + \alpha_t\big(h_\beta(Q_t) + M_{t+1}\big), \quad h_\beta(Q) = D(B + \beta I)(\mathcal{T}Q - Q), \quad \dot{Q} = h_\beta(Q), \tag{3}$$

where M_{t+1} denotes the Markovian noise induced by the trajectory $\{Y_t\}$. The term DB implements cross-action value sharing, whereas βD provides the diagonal residual correction. We show that the preconditioner matrix $D(B+\beta I)$ is invertible, thus ensuring that Q^* is the fixed point of the limiting ODE in Appendix B.2.

3.2 Asymptotic Guarantees

A contractive fixed-point representation. Our convergence analysis will use the Lyapunov framework for *contractive stochastic approximation* (Chen et al., 2024). To place equation 3 in this form, we introduce an auxiliary analysis parameter $\eta > 0$ and define the **V-SHARE fixed-point map**

$$F_{\eta,\beta}(Q) := Q + \eta h_\beta(Q). \tag{4}$$

Using Equation 4, Equation 3 can be rewritten as

$$Q_{t+1} = Q_t + \frac{\alpha_t}{\eta}\left[F_{\eta,\beta}(Q_t) - Q_t + \eta M_{t+1}\right]. \tag{5}$$

We show that $F_{\eta,\beta}$ is a contraction in the sup-norm (Lemma B.2 in Appendix B.2) whenever $\kappa_\beta := \min_{s,a} d(s)\big[(1-\gamma)\beta - (1+\gamma) + 2\pi_b(a|s)\big] > 0$ and $\eta \le L_\beta^{-1}$, $L_\beta := \max_{s,a} d(s)\big(\beta + \pi_b(a|s)\big)$. This contraction is used to establish the asymptotic convergence of the algorithm as follows.

Theorem 3.1 (Almost-sure convergence of V-SHARE under Markovian sampling)**.** *Under Assumption 1, let $\kappa_\beta > 0$, and $\alpha_t = \alpha_0/(t+h)^p$ for some $p \in (1/2, 1]$ and $\alpha_0, h > 0$. Then, the* V-SHARE *iterates (2) are almost surely bounded and satisfy*

$$Q_t \to Q^\star \qquad a.s.$$

Proof sketch. The full proof is given in Appendix B.3. Let $W(Q) := \|Q - Q^\star\|_\infty$. The contraction property established in Lemma B.2 implies that the upper-right Dini derivative (Khalil, 2002) of W along the limiting ODE $\dot{Q} = h_\beta(Q)$ satisfies

$$D^+W(Q; h_\beta(Q)) \le -\kappa_\beta W(Q).$$

Hence, $Q^\star$ is the unique globally exponentially stable equilibrium of the limiting ODE. Now, writing the V-SHARE recursion as $Q_{t+1} = Q_t + \alpha_t H_\beta(Q_t, Y_{t+1})$, where Y_t is the Markovian sample, we show that $H_\beta(Q, Y)$ is globally Lipschitz, its scaled limit exists, and the averaged limiting ODE has the origin as a globally exponentially stable equilibrium by the contraction argument of Lemma B.2. Assumption 1 verifies the required Markov averaging conditions of Liu et al. (2025). Liu et al. (2025, Theorem 7) therefore gives $\sup_t \|Q_t\|_\infty < \infty$ a.s. Their ODE convergence result (Liu et al., 2025, Corollary 8) then implies that the iterates approach a bounded invariant set of $\dot{Q} = h_\beta(Q)$. Since this ODE has the unique globally exponentially stable equilibrium $Q^\star$, the only bounded invariant set is $\{Q^\star\}$, and hence $Q_t \to Q^\star$ a.s.

Implications of Theorem 3.1. To our knowledge, Theorem 3.1 gives the first almost-sure convergence guarantee for an advantage-based algorithm under Markovian sampling. We show below that the same contractive regime also yields mean-square error bounds for the iterates, thereby providing finite-sample guarantees.

3.3 FINITE-TIME GUARANTEES

We now quantify the finite-sample behavior of V-SHARE under Markovian sampling. Recall the SA recursion equation 3 and suppose that $\kappa_\beta > 0$, so that the V-SHARE fixed-point map is contractive by Lemma B.2. We state the results for constant stepsizes here. The bounds for diminishing and decaying stepsizes are given in Theorem B.4 of Appendix B.4.

Theorem 3.2 (Finite-time convergence of V-SHARE)**.** *Under Assumption 1, suppose $\kappa_\beta > 0$. Let $\alpha_k \equiv \alpha$, α sufficiently small. Define, $K_{\mathrm{mix}} = \min\{k \ge 0 : k \ge t_{\alpha L_\beta}\}$,*

$$c_{2,\beta} := \left((2 + \frac{(1+\beta/b_{\min})(1+\gamma)}{L_\beta})\|Q^\star\|_\infty + \frac{(1+\beta/b_{\min})}{L_\beta}\right)^2.$$

Then, $\forall k \ge K_{\mathrm{mix}}$,

$$\mathbb{E}\|Q_k - Q^\star\|_\infty^2 \le \mathcal{O}\left(\|Q_0 - Q^*\|_\infty^2 \left(1 - \frac{\alpha\kappa_\beta}{2}\right)^{k - t_{\alpha L_\beta}} + \frac{\alpha L_\beta^3 \log|\mathcal{S}||\mathcal{A}|}{\kappa_\beta^2} c_{2,\beta}\, t_{\alpha L_\beta}\right). \tag{6}$$

Proof sketch. We provide the main steps here; the complete proof is given in Appendix B.4.

1. **Contractive-SA reduction.** With $\eta = L_\beta^{-1}$, Lemma B.2 gives a contraction factor $\rho = 1 - \kappa_\beta/L_\beta$ for the V-SHARE fixed point map. Hence the recursion is a contractive Markovian SA with effective stepsize $\epsilon_k = \alpha_k L_\beta$.
2. **Moreau-envelope drift under Markovian sampling.** Apply the generalized Moreau-envelope construction of Chen et al. (2024) to obtain a smooth Lyapunov function equivalent to $\|Q - Q^\star\|_\infty^2$. Contractivity yields a negative drift of order $\epsilon_k\kappa_\beta/L_\beta$, while temporal dependence is controlled over the mixing window, producing the Markovian error term proportional to $\epsilon_k\epsilon_{k-t_k,k-1}$ where $\alpha_{i,j} = \sum_{k=i}^{j}\alpha_k$.

3. **One-step recursion and unrolling.** Substituting the V-SHARE sample-map growth bounds into the Markovian Moreau recursion and imposing the mixing-window condition allows the correlation term to be absorbed into the negative drift. Unrolling the resulting scalar recursion for constant stepsizes yields equation 6.

Implications of Theorem 3.2. To our knowledge, Theorem 3.2 establishes the first finite-time convergence guarantees for an advantage-based algorithm under the challenging regime of Markovian sampling.

Remark 1 (Residual-preconditioning tradeoff)**.** *For large β, $L_\beta = \Theta(\beta)$ and $\kappa_\beta = \Theta(\beta)$, while $c_{2,\beta}$ remains bounded in β and scales as $O(b_{\min}^{-2})$. Hence the constant-stepsize stochastic term in Theorem 3.2 scales as (ignoring the dimensionality dependence)*

$$\nu_\beta := \frac{\alpha L_\beta^3 c_{2,\beta}}{\kappa_\beta^2} t_{\alpha L_\beta} = \Theta\big(\alpha\beta\, c_{2,\beta} t_{\alpha L_\beta}\big)\,.$$

Thus, larger β strengthens transient contraction but enlarges the constant-stepsize stochastic radius. We observe this effect in Figure 1a. This motivates the decreasing-β schedules of Section 3.4.

Remark 2 (Comparison with asynchronous Q-learning)**.** *Asynchronous Q-learning contracts at gap $\mu_{\min}(1-\gamma)$, where $\mu_{\min} := \min_{s,a} d(s)\pi_b(a|s)$, whereas for large β, $\kappa_\beta \sim d_{\min}(1-\gamma)\beta$. Thus* V-SHARE *removes the rare-action factor from the leading deterministic contraction gap and yields a faster contraction. For a detailed comparison, we defer the details to Appendix B.5.*

3.4 Decreasing Residual-Preconditioning Schedules

Remark 1 shows that stronger residual preconditioning improves transient contraction but enlarges the constant-stepsize stochastic radius ν_β. This suggests using a large value of β early in learning and decreasing it as the iterates approach $Q^\star$.

As a motivating example, consider the two-stage schedule

$$\beta_k = \begin{cases} \beta_h, & k < K, \\ \beta_\ell, & k \geq K, \end{cases} \qquad \beta_h \geq \beta_\ell, \qquad \kappa_{\beta_\ell} > 0.$$

The first stage contracts more aggressively, while the second stage reduces the stochastic error induced by residual preconditioning. Importantly, errors accumulated during the high-β phase are themselves contracted during the subsequent low-β phase.

In the following theorem, for notational ease we skip the subscript β in the various constants of Theorem 3.2 and instead write them with the subscripts h and l to denote β_h and β_l respectively. Let $K_{\text{mix}} = \min\{k \geq 0 : k \geq t_{\alpha L_h}\}$ and $K > K_{\text{mix}}$.

Theorem 3.3 (Two-stage residual preconditioning)**.** *Suppose the constant stepsize α satisfies the uniform Markovian mixing-window condition over $\beta \in [\beta_\ell, \beta_h]$. Then, $\forall T \geq K$,*

$$\mathbb{E}\|Q_T - Q^\star\|_\infty^2 = \mathcal{O}\left(\|Q_0 - Q^*\|_\infty^2 \exp\left\{-\frac{\alpha}{2}\left(\kappa_h(K - t_{\alpha L_h}) + \kappa_\ell(T - K - t_{\alpha L_l})\right)\right\}\right) + \mathcal{O}\left(\nu_h \exp\left\{-\frac{\alpha\kappa_\ell}{2}(T-K)\right\} + \nu_l\right).$$

Proof sketch. The full proof is given in Appendix C. We apply the fixed-β Markovian bound of Theorem 3.2 separately on the two stages. During the first K iterations, we use $\eta_h = L_h^{-1}$, yielding contraction at rate κ_h and a high-β stochastic radius ν_h.

At the switching time K, we restart the analysis conditionally on (Q_K, Y_K) with $\eta_\ell = L_\ell^{-1}$. Since the mixing-time bound is uniform over the initial Markov state, Theorem 3.2 applies to the second stage with the low-β constants $(L_\ell, \kappa_\ell, c_{2,\ell})$. Consequently, the error inherited from the first stage is further contracted by $\exp\{-\alpha\kappa_\ell(T-K)/2\}$, while noise generated after the switch contributes only the low-β stochastic radius ν_l. Combining the two stagewise bounds gives the result.

Implications of Theorem 3.3. The bound makes the benefit of scheduling explicit: large β can be used to accelerate the early transient, while the effect of its larger stochastic error is exponentially forgotten during the low-β tail. After a sufficiently long second stage, the dominant terminal term is ν_l. Thus, the low-β phase governs the terminal stochastic scale.

We derive explicit bounds for smoothly decreasing schedules such as linear, polynomial, and normalized-exponential in Appendix C. We compare the performance of the various schedules in Figure 1b.

3.5 TWO-TIMESCALE V-SHARE

The preceding single-timescale theory requires β to be large enough that the V-SHARE fixed point map is contractive (Lemma B.2). Two-timescale learning provides a complementary route that removes this restriction entirely. The key idea is to let the advantage component evolve faster than the shared value component. For every fixed $\beta > 0$, the fast advantage dynamics then reduce, after a simple change of variables, to standard Bellman dynamics with contraction gap $d_{\min}(1-\gamma)$.

Two-timescale V-SHARE (Algorithm 2) uses

$$V_{t+1}(S_t) = V_t(S_t) + \alpha_t \delta_t, \qquad A_{t+1}(S_t, a) = A_t(S_t, a) + \eta_t \frac{\beta\, \mathbf{1}\{U_t = a\}}{\pi_b(a|S_t)} \delta_t. \tag{7}$$

For fixed $\beta > 0$, define the **effective advantage stepsize** $\tau_t := \beta\eta_t$. We show the asymptotic convergence as follows:

Theorem 3.4 (Almost-sure convergence of two-timescale V-SHARE)**.** *Fix any $\beta > 0$ and let $\tau_t := \beta\eta_t$. Suppose Assumption 1 holds and*

$$\tau_t = \frac{\tau_0}{(t+h)^p}, \qquad \alpha_t = \frac{\alpha_0}{(t+h)^q}, \qquad \frac{1}{2} < p < q \le 1.$$

Then, $A_t - \big(Q^\star - LV_t\big) \to 0 \quad a.s. \implies Q_t = LV_t + A_t \to Q^\star \quad a.s.$

Proof sketch. The full proof is given in Appendix D.2. For a frozen value function V, the fast limiting ODE is $\dot{A} = D[\mathcal{T}(LV + A) - (LV + A)]$. Under the change of variables $Q = LV + A$, this becomes $\dot{Q} = D(\mathcal{T}Q - Q)$, whose unique globally exponentially stable equilibrium is $Q^\star$, with contraction gap $\kappa_{\text{fast}} = d_{\min}(1-\gamma)$. Hence the frozen fast recursion has equilibrium $\lambda(V) = Q^\star - LV$. The key observation is that the tracking error $A_t - \lambda(V_t) = Q_t - Q^\star$ evolves as a single-timescale Markovian stochastic approximation on the fast timescale, with the slow value update appearing only through the vanishing perturbation as $\alpha_t/\tau_t \to 0$. We therefore extend the Markovian ODE result of Liu et al. (2025) to stochastic approximation with a vanishing drift perturbation. We show that under their assumptions, stability and convergence to invariant sets of the unperturbed limiting ODE are preserved (Proposition D.1). This is the Markovian analogue of the vanishing-measurement-error extension in Borkar (2022, Section 2.2). Since the unperturbed limiting ODE has the unique equilibrium $Q^\star$, it follows that $A_t - \lambda(V_t) = Q_t - Q^\star \to 0$ a.s.

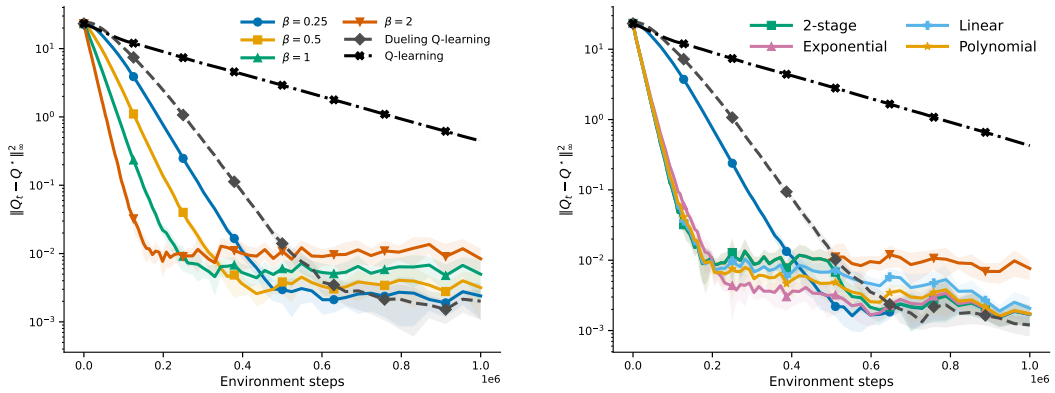
Implications of Theorem 3.4. First, Theorem 3.4 shows that timescale separation removes the single-timescale threshold on β: every fixed $\beta > 0$ is admissible once $\alpha_t/\tau_t \to 0$. *Second*, the proof establishes a Markovian analogue of the classical vanishing-measurement-error extension of stochastic approximation, showing that the stability and ODE limit theory of Liu et al. (2025) is preserved under a state-dependent drift perturbation whose coefficient vanishes in the limit.

Finite-time tracking. The asymptotic result shows that timescale separation eliminates the single-timescale threshold on β. We next quantify how quickly the fast advantage iterate tracks its moving equilibrium. Define $\Delta_t := A_t - A^\star(V_t) = Q_t - Q^\star$. The leading dynamics of Δ_t inherit the Bellman contraction gap $\kappa_{\text{fast}} = d_{\min}(1-\gamma)$, while motion of the value component acts only as a slower perturbation controlled by α_t/τ_t. Let $K := \min\{k : k \ge t_{\tau_k}\}$.

Theorem 3.5 (Finite-time tracking of two-timescale V-SHARE)**.** *Let Assumption 1 hold and the timescale separation satisfy $\frac{\alpha_t}{\tau_t} \le c_{sep} = \Theta(\frac{\kappa_{\text{fast}}}{2d_{\max}(1+\gamma)}) \quad \forall t$. For $\tau_t = \frac{\tau_0}{(t+h)^p}, \quad \alpha_t = \frac{\alpha_0}{(t+h)^q}, \quad \frac{1}{2} < p < q \le 1$, with appropriately chosen α_0, τ_0, h yields $\forall t > K$,*

$$\mathbb{E}\|Q_t - Q^\star\|_\infty^2 = \widetilde{\mathcal{O}}\left(\frac{1}{t^p}\right).$$

Proof sketch. The full proof is given in Appendix D.4. As in the asymptotic analysis, define $\Delta_t := A_t - (Q^\star - LV_t) = Q_t - Q^\star$. Its recursion is autonomous and can be written as $\Delta_{t+1} = \Delta_t +$



(a) Constant stepsize.

(b) Decreasing residual-preconditioning schedules.

Figure 1: **Tabular experiments on randomly generated MDPs.** (a): performance of single-timescale V-SHARE (Algorithm 1) for different residual weights β under constant stepsizes, compared with asynchronous Q-learning and Dueling Q-learning. (b): comparison of decreasing residual-preconditioning schedules that interpolate from β_{high} to β_{low}. Curves show the mean squared sup-norm error $\mathbb{E}\|Q_t - Q^\star\|_\infty^2$ over independent trajectories; shaded regions denote pointwise 95% confidence intervals. The y-axis is in the logarithmic scale in all plots.

$\tau_t \left[H_A(\Delta_t, Y_{t+1}) + \frac{\alpha_t}{\tau_t} H_V(\Delta_t, Y_{t+1}) \right]$. We view the advantage update as the dominant contractive Markovian SA and the value update as a state-dependent perturbation of size α_t/τ_t. The stationary mean operator of the fast recursion is contractive with gap $\kappa_{\text{fast}} = d_{\min}(1-\gamma)$, so we construct a fixed generalized Moreau-envelope Lyapunov function from this fast contraction. We then apply our extension of the Markovian finite-time analysis of Chen et al. (2024) (Section D.3). The perturbation contributes an additional mean-drift term, which is absorbed into the fast negative drift when α_t/τ_t is sufficiently small, and an additional Markovian correlation term of the same mixing-window order as the original sampling noise. This yields the same scalar Lyapunov recursion as in the single-timescale analysis, up to modified constants. Unrolling the recursion gives the desired rate.

Implications of Theorem 3.5. Theorem 3.5 establishes $\widetilde{\mathcal{O}}(1/t^p)$ mean square error rates for V-SHARE for all $\beta > 0$ removing the restriction of κ_β in the single-timescale setting. We evaluate the performance of two-timescale V-SHARE and compare it with the baselines of Asynchronous Q-learning and Dueling Q-learning in Figure 2b, Appendix F.1.

4 EXPERIMENTS

4.1 DISCRETE MDP

We first evaluate V-SHARE on tabular MDPs. Inspired by Tang et al. (2023), we consider a randomly generated finite MDP with 10 states and 5 actions. For each state-action pair (s, a), the transition distribution $P(\cdot \mid s, a)$ is sampled independently from a symmetric Dirichlet distribution, $P(\cdot \mid s, a) \sim \text{Dirichlet}(\alpha, \ldots, \alpha), \alpha = 0.5$. The rewards are sampled independently and uniformly from $[0, 1]$, and the discount factor is set to $\gamma = 0.95$. We compare V-SHARE and its variants against tabular Dueling Q-learning (Daley et al., 2025) and asynchronous Q-learning. Samples are generated by a fixed behavior policy $\pi_b = (1-\epsilon)\pi_{\text{det}} + \epsilon u$ where π_{det} is a randomly sampled deterministic policy, u is the uniform policy over actions, and $\epsilon = 0.8$. The MDP and behavior policy are fixed across all repetitions, while the sample trajectory is varied across 10 independent seeds. We observe that under constant stepsizes, increasing β substantially accelerates the initial decay but eventually leads to a larger error floor (Figure 1a), illustrating the predicted contraction-variance tradeoff. Decreasing-β schedules combine these effects, matching the rapid initial progress of strong residual preconditioning while attaining a smaller terminal error (Figure 1b). In all the plots, across all settings, V-SHARE converges faster than asynchronous Q-learning and Dueling Q-learning.

Environment	DQN	ME Layer	Soft RDQ	V-SHARE LR Multiplier					V-SHARE β-Scheduling
				0.25×	0.5×	1×	2×	4×	
Acrobot-v1	-443.12 ± 101.39	-140.40 ± 113.25	-85.44 ± 9.30	$\mathbf{-82.44 \pm 11.32}$	-103.80 ± 39.53	-87.24 ± 4.65	-169.08 ± 185.00	-167.24 ± 186.44	-108.72 ± 48.35
CartPole-v1	375.08 ± 143.63	342.40 ± 195.66	423.24 ± 171.64	$\mathbf{500.00 \pm 0.00}$	$\mathbf{500.00 \pm 0.00}$	$\mathbf{500.00 \pm 0.00}$	464.00 ± 77.52	$\mathbf{500.00 \pm 0.00}$	$\mathbf{500.00 \pm 0.00}$
LunarLander-v2	218.37 ± 36.83	193.98 ± 101.57	244.94 ± 25.36	$\mathbf{248.79 \pm 11.53}$	227.80 ± 51.22	202.93 ± 4.76	105.49 ± 98.36	-24.65 ± 76.29	168.55 ± 90.31
MountainCar-v0	-166.20 ± 46.36	-138.04 ± 16.81	$\mathbf{-106.24 \pm 5.87}$	-183.36 ± 37.21	-136.92 ± 9.06	-131.48 ± 22.14	-133.60 ± 17.56	-141.40 ± 35.89	-129.28 ± 4.98
Asterix-v1	42.33 ± 14.28	41.90 ± 12.77	$\mathbf{55.97 \pm 9.13}$	49.57 ± 10.05	41.47 ± 15.00	47.97 ± 20.52	35.70 ± 7.52	27.03 ± 4.82	41.53 ± 11.08
Breakout-v1	31.33 ± 13.69	39.83 ± 15.19	37.87 ± 16.67	30.17 ± 7.69	34.07 ± 12.29	$\mathbf{53.67 \pm 20.50}$	40.63 ± 14.13	33.93 ± 12.58	40.20 ± 15.59
Freeway-v1	61.77 ± 1.50	61.40 ± 1.48	$\mathbf{63.10 \pm 1.22}$	61.77 ± 2.77	61.67 ± 1.82	59.90 ± 2.27	61.30 ± 3.14	55.37 ± 7.19	60.17 ± 2.06
Seaquest-v1	74.73 ± 22.11	119.00 ± 32.18	122.40 ± 46.38	$\mathbf{132.07 \pm 53.50}$	91.37 ± 30.00	103.03 ± 39.83	80.03 ± 28.40	61.43 ± 10.66	114.70 ± 32.40
SpaceInvaders-v1	196.57 ± 43.51	173.83 ± 68.47	254.43 ± 106.39	222.67 ± 41.82	$\mathbf{282.20 \pm 59.04}$	192.70 ± 91.94	136.67 ± 66.44	134.57 ± 67.59	134.33 ± 68.78

Table 1: **Deep RL results on Classic Control and MinAtar Environments.** We compare different configuration of V-SHARE against DQN (Mnih et al., 2015), ME Layer (Nagarajan et al., 2026), and Soft RDQ (Daley et al., 2025) across four Classic Control and five MinAtar environments. Results report the mean $\pm$ standard deviation over five independent training seeds. The results show that while it is hard to directly apply the theoretical results to the deep RL setting, V-SHARE tends to perform at par with Soft RDQ and significantly outperforms vanilla DQN.

4.2 DEEP RL

To verify whether our mechanism suggested by the tabular analysis remains effective in a more complex setting, we evaluate our method on both Classic Control and MinAtar environments. Specifically, we study three settings of V-SHARE: single-timescale learning, two-timescale learning, and β scheduling. In the *single-timescale* setting, the value and advantage networks use the same learning rate. In the *two-timescale* setting, we vary the advantage learning rate relative to the value learning rate, motivated by our analysis showing that timescale separation can remove the large-β requirement. In the β scheduling setting, the learning rates are kept equal while β is linearly decayed from 100 to 1, motivated by the contraction-variance tradeoff.

Table 1 summarizes our deep RL results across both MinAtar and Classic Control environments. Overall, V-SHARE is competitive with the considered baselines, but its performance is sensitive to the relative learning rates of the value and advantage components. No single advantage learning-rate multiplier performs best across all tasks: smaller multipliers are preferred in several environments, while intermediate settings perform better in others. In particular, the best-performing V-SHARE configuration achieves the highest mean return on six of the nine environments. At the same time, larger advantage learning-rate multipliers such as $2\times$ and $4\times$ can substantially degrade performance on some tasks. These results indicate that timescale separation can meaningfully affect deep RL, but the optimal relative update speed of the value and advantage components is environment dependent.

5 DISCUSSION AND FUTURE WORK

Key take-aways. We introduce V-SHARE, an advantage-based RL algorithm, and provide the first asymptotic and non-asymptotic convergence guarantees for an advantage-based method under Markovian sampling. Our results suggest several practical principles for advantage-based value learning. **First**, they support separating the shared state-value component from action-specific advantages. **Second**, the residual weight β controls how aggressively the algorithm corrects infrequently sampled actions, motivating stronger advantage updates early and weaker ones later. **Third**, when low-advantage actions are also sampled infrequently, applying a higher effective learning rate to them can improve learning. **Finally**, the two-timescale analysis supports using a larger effective learning rate for advantages than for state values in tabular MDPs, while our deep RL experiments suggest that reducing the advantage learning rate can improve stability by limiting noise.

Future work. Several directions remain open. First, while there is supporting intuition, our analysis does not yet explain the empirical acceleration of dueling Q-learning over standard Q-learning. A related question is whether value sharing can yield sharper problem-dependent or sample-complexity guarantees. Second, extending the analysis of V-SHARE to function approximation is an important direction. Finally, our analysis assumes a fixed behavior policy; extending it to iterate-dependent sampling policies, such as ϵ-greedy exploration, is another natural direction.

AI USE STATEMENT

In this work, we used generative AI tools solely for proofreading the manuscript and for suggestions to improve the writing style and clarity.

REFERENCES

Jinane Abounadi, Dimitri P. Bertsekas, and Vivek S. Borkar. Stochastic Approximation for Nonexpansive Maps: Application to Q-Learning Algorithms. *SIAM Journal on Control and Optimization*, 41(1), 2002. 20

Leemon C. Baird. Advantage Updating. Technical Report WL-TR-93-1146, Wright-Patterson Air Force Base, 1993. 2

V. S. Borkar and S. P. Meyn. The O.D.E. Method for Convergence of Stochastic Approximation and Reinforcement Learning. *SIAM Journal on Control and Optimization*, 38(2), 2000. 2, 17

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Hindustan Book Agency, 2 edition, 2022. 3, 7, 31

Siddharth Chandak, Shaan Ul Haque, and Nicholas Bambos. Finite-time bounds for two-time-scale stochastic approximation with arbitrary norm contractions and markovian noise. In *IEEE Conference on Decision and Control*, 2025. 47

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-Sample Analysis of Contractive Stochastic Approximation Using Smooth Convex Envelopes. In *Advances in Neural Information Processing Systems*, volume 33, 2020. 2, 48

Zaiwei Chen, Siva T. Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. A Lyapunov Theory for Finite-Sample Guarantees of Markovian Stochastic Approximation. *Operations Research*, 72(4), 2024. 2, 4, 5, 8, 20, 21, 22, 23, 24, 25, 26, 27, 38, 39, 40, 41, 44, 48

Brett Daley, Prabhat Nagarajan, Martha White, and Marlos C. Machado. An Analysis of Action-Value Temporal-Difference Methods That Learn State Values. *Reinforcement Learning Journal*, 6, 2025. 2, 3, 8, 9, 14, 50

Abhijit Gosavi. Boundedness of Iterates in Q-Learning. *Systems & Control Letters*, 55(4), 2006. 20

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2012. 16

Shengyi Huang, Rousslan Fernand Julien Dossa, Chang Ye, Jeff Braga, Dipam Chakraborty, Kinal Mehta, and João G.M. Araújo. CleanRL: High-quality Single-file Implementations of Deep Reinforcement Learning Algorithms. *Journal of Machine Learning Research*, 23(274), 2022. 49

Tommi Jaakkola, Michael I. Jordan, and Satinder P. Singh. Convergence of Stochastic Iterative Dynamic Programming Algorithms. In *Advances in Neural Information Processing Systems*, volume 6, 1993. 1

Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, 3 edition, 2002. 5, 19, 30, 37

Donghwan Lee. Spectral Analysis of Dueling Q-Learning. *arXiv preprint arXiv:2607.08340*, 2026. 2

Shuze Daniel Liu, Shuhang Chen, and Shangtong Zhang. The ODE Method for Stochastic Approximation and Reinforcement Learning with Markovian Noise. *Journal of Machine Learning Research*, 26(24), 2025. 2, 5, 7, 17, 18, 19, 20, 31, 32, 33, 34, 35, 36

Vagul Mahadevan, Claire Chen, Shuze Daniel Liu, and Shangtong Zhang. Convergence of two-timescale markovian stochastic approximations with applications in reinforcement learning. In *International Conference of Machine Learning*, 2026. 31

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-Level Control through Deep Reinforcement Learning. *Nature*, 518(7540), 2015. 1, 9

Arsenii Mustafin, Aleksei Pakharev, Alex Olshevsky, and Ioannis Paschalidis. MDP Geometry, Normalization and Reward Balancing Solvers. In *International Conference on Artificial Intelligence and Statistics*, 2025. 2

Prabhat Nagarajan, Brett Daley, Martha White, and Marlos C. Machado. Accelerating Q-Learning through Efficient Value-Sharing across Actions. In *International Conference on Machine Learning*, 2026. 1, 2, 9

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-Dimensional Continuous Control Using Generalized Advantage Estimation. *arXiv preprint arXiv:1506.02438*, 2015. 2

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal Policy Optimization Algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998. 1

Yunhao Tang, Rémi Munos, Mark Rowland, and Michal Valko. VA-Learning as a More Efficient Alternative to Q-Learning. In *International Conference on Machine Learning*, 2023. 2, 8

John N. Tsitsiklis. Asynchronous Stochastic Approximation and Q-Learning. *Machine Learning*, 16(3), 1994. 1

Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas. Dueling Network Architectures for Deep Reinforcement Learning. In *International Conference on Machine Learning*, 2016. 1, 2, 50

Christopher J. C. H. Watkins and Peter Dayan. Q-Learning. *Machine Learning*, 8(3), 1992. 1

Kenny Young and Tian Tian. MinAtar: An Atari-Inspired Testbed for Thorough and Reproducible Reinforcement Learning Experiments. *arXiv preprint arXiv:1903.03176*, 2019. 2

Appendix

Table of Contents

A Background Derivations **13**
- A.1 Sampling model and Bellman residual . . . 13
- A.2 Asynchronous Q-learning . . . 14
  - A.2.1 Sup-norm contraction . . . 14
- A.3 Value-advantage update and induced Q-dynamics . . . 14
- A.4 Diagonal versus cross-action preconditioning . . . 15

B Single-Timescale V-SHARE: Detailed Analysis **16**
- B.1 V-SHARE recursion and stationary mean operator . . . 16
- B.2 Fixed-point preservation and contraction . . . 16
- B.3 Almost-sure convergence under Markovian sampling . . . 17
- B.4 Finite-time analysis under Markovian sampling . . . 20
- B.5 Dependence on β and comparison with Q-learning . . . 24

C Decreasing Residual-Preconditioning Schedules **25**
- C.1 Two-stage schedule . . . 25
- C.2 General decreasing schedules . . . 26
- C.3 Smoothly decreasing schedules . . . 28

D Two-Timescale V-SHARE: Detailed Proofs **29**
- D.1 Stochastic-approximation form and frozen fast dynamics . . . 29
- D.2 Almost-sure tracking under timescale separation . . . 31
  - D.2.1 Markovian stochastic approximation with vanishing perturbations . . . 31
  - D.2.2 Specialization to two-timescale V-SHARE . . . 35
- D.3 Finite-time Markovian stochastic approximation with a small perturbation . . . 38
- D.4 Finite-time tracking under timescale separation . . . 44

E Useful Technical Lemmas **47**

F Experimental Details **48**
- F.1 Discrete MDP . . . 48
- F.2 Deep RL Experiments . . . 48

Overview of the Appendix

- **Appendix A: Background and operator preliminaries.** Reviews asynchronous Q-learning and the value-advantage decomposition, derives the induced Q-dynamics of value sharing, and contrasts diagonal Q-learning updates with the cross-action preconditioning structure that motivates V-SHARE.
- **Appendix B: Single-timescale V-SHARE: detailed analysis.** Develops the single-timescale stochastic-approximation representation, proves fixed-point preservation and the contraction regime, establishes almost-sure convergence under Markovian sampling, and derives the finite-time bounds for constant and diminishing stepsizes. It also studies the dependence on β and compares the resulting guarantees with asynchronous Q-learning.
- **Appendix C: Decreasing residual-preconditioning schedules.** Analyzes time-varying choices of β. We first prove a sharp two-stage result showing how strong early contraction can be combined with a low-β terminal noise floor, and then derive conservative guarantees for general decreasing schedules and their linear, polynomial, and exponential specializations.
- **Appendix D: Two-timescale V-SHARE: detailed proofs.** Develops the two-timescale stochastic-approximation formulation and the frozen-fast dynamics, proves almost-sure tracking for every fixed $\beta > 0$, and derives finite-time tracking guarantees. The analysis also establishes auxiliary results for Markovian stochastic approximation with vanishing drift perturbations.
- **Appendix E: Useful Technical Lemmas.** Collects technical lemmas about the Moreau envelope construction used in the finite-sample analysis
- **Appendix F: Experimental details.** Provides additional discrete-MDP experiments, implementation details, hyperparameter settings, and the deep-RL experimental setup used to study single-timescale, decreasing-β, and two-timescale variants of V-SHARE.

A Background Derivations

This appendix derives the sample recursions and stationary mean fields used in Section 2, and recalls the standard sup-norm contraction argument for asynchronous Q-learning.

A.1 Sampling model and Bellman residual

Let $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$ be a finite discounted MDP with $\gamma \in [0, 1)$ and $|r(s, a)| \leq 1$. The Bellman optimality operator is

$$(\mathcal{T}Q)(s, a) = r(s, a) + \gamma \sum_{s'} P(s'|s, a) \max_{a'} Q(s', a').$$

It is a γ-contraction in $\|\cdot\|_\infty$ and therefore has a unique fixed point $Q^\star$.

Let $\pi_b(\cdot|s)$ be the behavior policy from Assumption 1. The induced state chain is irreducible and aperiodic with invariant distribution d, and

$$Y_{t+1} := (s_t, a_t, s_{t+1})$$

is a finite-state irreducible and aperiodic Markov chain with invariant distribution

$$d_Y(s, a, s') = d(s)\pi_b(a|s)P(s'|s, a).$$

For $y = (s, a, s')$, define

$$\delta(Q, y) := r(s, a) + \gamma \max_{a'} Q(s', a') - Q(s, a).$$

Its stationary conditional average over the successor state is

$$\sum_{s'} P(s'|s, a)\delta(Q, (s, a, s')) = (\mathcal{T}Q - Q)(s, a).$$

A.2 Asynchronous Q-learning

Asynchronous Q-learning updates only the sampled state-action coordinate:

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t \mathbf{1}\{s_t = s, a_t = a\}\delta(Q_t, Y_{t+1}).$$

Define the sample field

$$H_{\mathrm{Q}}(Q,y)(s_0,a_0) := \mathbf{1}\{s_0 = s, a_0 = a\}\delta(Q,y), \qquad y = (s,a,s').$$

Then

$$Q_{t+1} = Q_t + \alpha_t H_{\mathrm{Q}}(Q_t, Y_{t+1}).$$

Its stationary mean operator is

$$\begin{aligned}\bar{H}_{\mathrm{Q}}(Q)(s,a) &:= \mathbb{E}_{Y\sim d_Y}[H_{\mathrm{Q}}(Q,Y)(s,a)] \\ &= d(s)\pi_b(a|s)\sum_{s'} P(s'|s,a)\delta(Q,(s,a,s')) \\ &= d(s)\pi_b(a|s)(TQ - Q)(s,a).\end{aligned}$$

Thus, with

$$D_\mu := \mathrm{diag}\{d(s)\pi_b(a|s) : (s,a) \in \mathcal{S}\times\mathcal{A}\},$$

the stationary mean operator and associated limiting ODE are

$$\bar{H}_{\mathrm{Q}}(Q) = D_\mu(TQ - Q), \qquad \dot{Q} = D_\mu(TQ - Q).$$

A.2.1 Sup-norm contraction

Define

$$F_{\mathrm{Q}}(Q) := Q + D_\mu(TQ - Q).$$

For $i = (s,a)$, write $\mu_i := d(s)\pi_b(a|s)$. Then

$$[F_{\mathrm{Q}}(Q) - F_{\mathrm{Q}}(Q')]_i = (1-\mu_i)(Q_i - Q'_i) + \mu_i[(\mathcal{T}Q)_i - (\mathcal{T}Q')_i],$$

and hence

$$|[F_{\mathrm{Q}}(Q) - F_{\mathrm{Q}}(Q')]_i| \le [1 - \mu_i(1-\gamma)]\|Q - Q'\|_\infty.$$

Therefore

$$\|F_{\mathrm{Q}}(Q) - F_{\mathrm{Q}}(Q')\|_\infty \le [1 - \mu_{\min}(1-\gamma)]\|Q - Q'\|_\infty, \tag{8}$$

where

$$\mu_{\min} := \min_{s,a} d(s)\pi_b(a|s).$$

Thus the contraction gap is $\mu_{\min}(1-\gamma)$. The argument works because D_μ is diagonal: every coordinate is a convex combination of its previous value and its own Bellman update.

A.3 Value-advantage update and induced Q-dynamics

Consider the decomposition

$$Q_t(s,a) = V_t(s) + A_t(s,a)$$

and the hard RDQ update (Daley et al., 2025)

$$V_{t+1}(s_t) = V_t(s_t) + \alpha_t\delta(Q_t, Y_{t+1}), \quad A_{t+1}(s_t,a) = A_t(s_t,a) + \alpha_t\mathbf{1}\{a_t = a\}\delta(Q_t, Y_{t+1}),$$

with all coordinates at states $s \neq s_t$ unchanged. The decomposition is not identifiable: for any state-dependent scalar $c(s)$,

$$V_t(s) \mapsto V_t(s) + c(s), \qquad A_t(s,a) \mapsto A_t(s,a) - c(s)$$

leaves Q_t unchanged. We therefore analyze the iterate Q_t.

For every coordinate (s,a),

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t\mathbf{1}\{s_t = s\}[1 + \mathbf{1}\{a_t = a\}]\delta(Q_t, Y_{t+1}).$$

The first term is the value increment shared across all actions at the sampled state; the second is the sampled-action advantage increment.

For each state s, let

$$b_s := (\pi_b(a|s))_{a\in\mathcal{A}}, \qquad B_s := \mathbf{e}b_s^\top, \qquad B := \mathrm{blkdiag}(B_s : s \in \mathcal{S}),$$

and define the lifted state-visitation matrix

$$D := \mathrm{diag}\{d(s) : (s,a) \in \mathcal{S}\times\mathcal{A}\}.$$

The stationary mean operator of the shared value update is

$$DB(TQ - Q),$$

because for every (s,a),

$$\mathbb{E}_{Y\sim d_Y}\left[\mathbf{1}\{S=s\}\delta(Q,Y)\right] = d(s)\sum_c \pi_b(c|s)(TQ-Q)(s,c).$$

The sampled-action contribution is

$$D_\mu(TQ - Q).$$

Hence the stationary mean operator of the hard value-advantage update is

$$\bar{H}_{\mathrm{AV}}(Q) = (DB + D_\mu)(TQ - Q),$$

with associated limiting ODE

$$\dot{Q} = (DB + D_\mu)(TQ - Q).$$

For a 2-state, 2-action MDP, the preconditioners look like the following: Let,

$$\pi_b(i,j) := \pi_b(a_j \mid s_i), \quad d_i = d(s_i).$$

where $d(\cdot)$ is the stationary distirbution induced by the behavior policy $\pi_b(\cdot|\cdot)$

$$D_\mu = \begin{bmatrix} d_1\pi_b(1,1) & 0 & 0 & 0 \\ 0 & d_1\pi_b(1,2) & 0 & 0 \\ 0 & 0 & d_2\pi_b(2,1) & 0 \\ 0 & 0 & 0 & d_2\pi_b(2,2) \end{bmatrix}.$$

$$DB = D_\mu + \begin{bmatrix} d_1\pi_b(1,1) & d_1\pi_b(1,2) & 0 & 0 \\ d_1\pi_b(1,1) & d_1\pi_b(1,2) & 0 & 0 \\ 0 & 0 & d_2\pi_b(2,1) & d_2\pi_b(2,2) \\ 0 & 0 & d_2\pi_b(2,1) & d_2\pi_b(2,2) \end{bmatrix}.$$

A.4 Diagonal versus cross-action preconditioning

The preceding derivations give

$$\bar{H}_{\mathrm{Q}}(Q) = D_\mu(TQ - Q)$$

for asynchronous Q-learning and

$$\bar{H}_{\mathrm{AV}}(Q) = (DB + D_\mu)(TQ - Q)$$

for the hard RDQ update.

The Q-learning preconditioner D_μ is diagonal. In contrast, the shared value component satisfies

$$[DB]_{(s,a),(s,c)} = d(s)\pi_b(c|s)$$

for every pair of actions a, c at the same state. Thus a Bellman residual associated with action c contributes to the mean update of every action at state s.

For the hard RDQ preconditioner,

$$[DB + D_\mu]_{(s,a),(s,c)} = d(s)\pi_b(c|s) + \mathbf{1}\{a=c\}d(s)\pi_b(a|s).$$

In particular,

$$[DB + D_\mu]_{(s,a),(s,a)} = 2d(s)\pi_b(a|s),$$

while for $c \neq a$,

$$[DB + D_\mu]_{(s,a),(s,c)} = d(s)\pi_b(c|s) > 0.$$

This cross-action coupling is precisely what prevents the coordinatewise convex-combination argument used for Q-learning (equation 8) from carrying over.

B SINGLE-TIMESCALE V-SHARE: DETAILED ANALYSIS

This appendix proves the single-timescale results stated in Sections 3.1 and 3.3. Throughout, sampling is generated by the Markov chain $\{Y_t\}$ above. Recall $B_s = \mathbf{e}b_s^\top$, $B = \mathrm{blkdiag}(B_s : s \in \mathcal{S})$, and $D_{(s,a),(s,a)} = d(s)$.

B.1 V-SHARE RECURSION AND STATIONARY MEAN OPERATOR

The single-timescale V-SHARE updates are

$$V_{t+1}(s_t) = V_t(s_t) + \alpha_t \delta(Q_t, Y_{t+1}), \quad A_{t+1}(s_t, a) = A_t(s_t, a) + \alpha_t \frac{\beta \mathbf{1}\{a_t = a\}}{\pi_b(a|s_t)} \delta(Q_t, Y_{t+1}),$$

where $Q_t = LV_t + A_t$. Therefore the action-value iterate obeys

$$Q_{t+1}(s,a) = Q_t(s,a) + \alpha_t \mathbf{1}\{s_t = s\} \left[1 + \beta \frac{\mathbf{1}\{a_t = a\}}{\pi_b(a|s)}\right] \delta(Q_t, Y_{t+1}).$$

For $y = (s, a, s')$, define the sample field

$$H_\beta(Q, y)(s_0, a_0) := \mathbf{1}\{s_0 = s\} \left[1 + \beta \frac{\mathbf{1}\{a_0 = a\}}{\pi_b(a|s)}\right] \delta(Q, y).$$

Then the recursion is

$$Q_{t+1} = Q_t + \alpha_t H_\beta(Q_t, Y_{t+1}).$$

Its stationary mean operator is

$$h_\beta(Q) := \mathbb{E}_{Y \sim d_Y}[H_\beta(Q, Y)].$$

For a fixed coordinate (s, a), the shared-value contribution is

$$\mathbb{E}_{Y \sim d_Y}\left[\mathbf{1}\{S = s\}\delta(Q, Y)\right] = d(s) \sum_u b(u|s)(TQ - Q)(s, u) = [DB(TQ - Q)](s, a),$$

while importance weighting gives

$$\mathbb{E}_{Y \sim d_Y}\left[\mathbf{1}\{S = s\} \frac{\mathbf{1}\{A = a\}}{\pi_b(a|s)} \delta(Q, Y)\right] = d(s)(TQ - Q)(s, a).$$

Consequently,

$$h_\beta(Q) = D(B + \beta I)(TQ - Q).$$

The associated limiting ODE is

$$\dot{Q} = D(B + \beta I)(\mathcal{T}Q - Q).$$

For later use, write $G_\beta := D(B + \beta I)$. For $i = (s, a)$,

$$(G_\beta)_{ii} = d(s)[\beta + \pi_b(a|s)], \quad \sum_{j \neq i} (G_\beta)_{ij} = d(s)[1 - \pi_b(a|s)], \quad \sum_j (G_\beta)_{ij} = d(s)(\beta + 1). \tag{9}$$

Thus residual preconditioning leaves the cross-action sharing term DB unchanged while adding a diagonal correction βD.

B.2 FIXED-POINT PRESERVATION AND CONTRACTION

We first verify that residual preconditioning preserves the Bellman fixed point.

Lemma B.1 (Fixed-point preservation)**.** *For every $\beta > 0$, $Q^\star$ is the unique equilibrium of $\dot{Q} = h_\beta(Q)$.*

Proof. For every state s, the matrix $\beta I + \mathbf{e}b_s^\top$, where I is the identity matrix in $\mathbb{R}^{|\mathcal{A}| \times |\mathcal{A}|}$, is invertible since the matrix determinant lemma (Horn & Johnson, 2012) gives

$$\det(\beta I + \mathbf{e}b_s^\top) = \beta^{|\mathcal{A}|-1}(\beta + 1) > 0.$$

Since D is also invertible, $D(B + \beta I)(\mathcal{T}Q - Q) = 0$ iff $\mathcal{T}Q = Q$. The Bellman optimality operator has the unique fixed point $Q^\star$. □

We next prove the contraction result used throughout the paper. Define

$$F_{\eta,\beta}(Q) := Q + \eta h_\beta(Q),$$

and recall

$$L_\beta := \max_{s,a} d(s)\big(\beta + \pi_b(a|s)\big), \qquad \kappa_\beta := \min_{s,a} d(s)\left[(1-\gamma)\beta - (1+\gamma) + 2b(a|s)\right]. \tag{10}$$

Lemma B.2 (Contraction of the V-SHARE fixed-point map)**.** *Suppose $\kappa_\beta > 0$. For every $0 < \eta \le L_\beta^{-1}$,*

$$\|F_{\eta,\beta}(Q) - F_{\eta,\beta}(Q')\|_\infty \le (1 - \eta\kappa_\beta)\|Q - Q'\|_\infty.$$

Consequently, it suffices that

$$\beta > \max\left\{0,\ \frac{1+\gamma-2b_{\min}}{1-\gamma}\right\}.$$

Proof. Let $z := Q - Q'$, $M := \|z\|_\infty$, and $\Delta_T := \mathcal{T}Q - \mathcal{T}Q'$. The Bellman contraction gives $\|\Delta_T\|_\infty \le \gamma M$.

For $i = (s, a)$,

$$[F_{\eta,\beta}(Q) - F_{\eta,\beta}(Q')]_i = \big(1 - \eta(D(B+\beta I))_{ii}\big)z_i - \eta\sum_{j\neq i}(D(B+\beta I))_{ij}z_j + \eta\sum_j (D(B+\beta I))_{ij}(\Delta_T)_j.$$

Since $\eta \le L_\beta^{-1}$, $1 - \eta(D(B+\beta I))_{ii} \ge 0$. Using equation 9,

$$\begin{aligned}
|[F_{\eta,\beta}(Q) - F_{\eta,\beta}(Q')]_i| &\le \Big[1 - \eta d(s)(\beta + \pi_b(a|s)) + \eta d(s)(1 - \pi_b(a|s)) + \eta\gamma d(s)(\beta+1)\Big]M \\
&= \Big[1 - \eta d(s)\big\{(1-\gamma)\beta - (1+\gamma) + 2b(a|s)\big\}\Big]M \\
&\le (1 - \eta\kappa_\beta)M.
\end{aligned}$$

Taking the maximum over i proves the contraction claim. Positivity of κ_β follows whenever $(1-\gamma)\beta - (1+\gamma) + 2b(a|s) > 0$ for every (s, a). □

B.3 Almost-sure convergence under Markovian sampling

We prove almost-sure convergence of V-SHARE directly under the Markovian trajectory generated by the behavior policy, using the stochastic-approximation (SA) framework of Liu et al. (2025), which is an extension of the Borkar-Meyn theorem (Borkar & Meyn, 2000) to the Markovian noise setting. Their recursion has the form $Q_{t+1} = Q_t + \alpha_t H(Q_t, Y_{t+1})$. We verify Liu et al. (2025, Assumptions 1-6) in order to apply Liu et al. (2025, Theorem 7, Corollary 8).

Markovian SA representation. Let $Y_{t+1} := (s_t, a_t, s_{t+1})$. For $y = (s, a, s')$, define $\delta(Q, y) := r(s, a) + \gamma \max_{a'} Q(s', a') - Q(s, a)$ and

$$H_\beta(Q, y)(s_0, a_0) := \mathbf{1}\{s_0 = s\}\left(1 + \beta\frac{\mathbf{1}\{a = a_0\}}{\pi_b(a|s)}\right)\delta(Q, y).$$

Then the V-SHARE recursion is

$$Q_{t+1} = Q_t + \alpha_t H_\beta(Q_t, Y_{t+1}). \tag{11}$$

Under stationarity, $\mathbb{P}(Y = (s, a, s')) = d(s)\pi_b(a|s)P(s'|s, a)$, and therefore

$$h_\beta(Q) := \mathbb{E}_{Y\sim d_Y}[H_\beta(Q, Y)] = D(B + \beta I)(\mathcal{T}Q - Q).$$

Assumption 1: unique invariant distribution. Under Assumption 1, the state chain induced by π_b is finite, irreducible, and aperiodic. Since π_b has full support, the transition chain $Y_t = (s_{t-1}, a_{t-1}, s_t)$ is finite-state and ergodic on its communicating class, with unique invariant distribution $d_Y(s, u, s') = d(s)\pi_b(u|s)P(s'|s, u)$. Thus Assumption 1 holds.

Assumption 2: learning-rate regularity. We take

$$\alpha_t = \frac{\alpha_0}{(t+h)^p}, \qquad \frac{1}{2} < p \le 1.$$

Then $\alpha_t > 0$, $\alpha_t \downarrow 0$, and $\sum_t \alpha_t = \infty$.

Liu et al. (2025, Remark 1) implies that Assumption 2 holds.

Assumption 3: scaled-update limit. For $c \geq 1$, define $H_{\beta,c}(Q, y) := H_\beta(cQ, y)/c$ and $y = (s, a, s')$. Then

$$H_{\beta,c}(Q, y)(s_0, a_0) = \mathbf{1}\{s_0 = s\} \left(1 + \beta \frac{\mathbf{1}\{a_0 = a\}}{\pi_b(a|s)}\right) \left[\frac{r(s, a)}{c} + \gamma \max_{a'} Q(s', a') - Q(s, a)\right].$$

Define

$$H_{\beta,\infty}(Q, y)(s_0, a_0) := \mathbf{1}\{s_0 = s\} \left(1 + \beta \frac{\mathbf{1}\{a_0 = a\}}{\pi_b(a|s)}\right) \left[\gamma \max_{a'} Q(s', a') - Q(s, a)\right].$$

Hence

$$H_{\beta,c}(Q, y) - H_{\beta,\infty}(Q, y) = \frac{1}{c} R_\beta(y), \tag{12}$$

where

$$R_\beta(y)(s_0, a_0) := \mathbf{1}\{s_0 = s\} \left(1 + \beta \frac{\mathbf{1}\{a_0 = a\}}{\pi_b(a|s)}\right) r(s, a).$$

Thus Assumption 3 holds with $\kappa(c) = 1/c$ and $b(Q, y) = R_\beta(y)$. Since $b(Q, y)$ is independent of Q, its Lipschitz constant can be taken as $L_b(y) \equiv 0$.

Assumption 4: Lipschitz continuity and integrability. For any Q, Q' and $y = (s, a, s')$,

$$|\delta(Q, y) - \delta(Q', y)| \leq (1 + \gamma)\|Q - Q'\|_\infty.$$

Therefore, with $b_{\min} := \min_{s,a} \pi_b(a|s) > 0$,

$$\|H_\beta(Q, y) - H_\beta(Q', y)\|_\infty \leq \left(1 + \frac{\beta}{b_{\min}}\right)(1 + \gamma)\|Q - Q'\|_\infty.$$

The same bound holds for $H_{\beta,\infty}$. Hence Assumption 4 holds with the constant Lipschitz function $L(y) \equiv (1 + \beta/b_{\min})(1 + \gamma)$, whose invariant expectation is finite. Since the Markov state space is finite and rewards are bounded, the remaining required integrability conditions also hold.

Averaging $H_{\beta,\infty}$ gives

$$h_{\beta,\infty}(Q) = D(B + \beta I)(\mathcal{T}_0 Q - Q), \tag{13}$$

where $(\mathcal{T}_0 Q)(s, a) := \gamma \sum_{s'} P(s'|s, a) \max_{a'} Q(s', a')$.

Assumption 5: scaling limit and stability at infinity. Let $h_{\beta,c}(Q) := h_\beta(cQ)/c$. Taking expectations in equation 12 gives

$$h_{\beta,c}(Q) - h_{\beta,\infty}(Q) = \frac{1}{c} \mathbb{E}_{d_Y}[R_\beta(Y)].$$

Hence, for every compact set K,

$$\sup_{Q \in K} \|h_{\beta,c}(Q) - h_{\beta,\infty}(Q)\|_\infty \leq \frac{C_R}{c} \to 0,$$

where $C_R := \|\mathbb{E}_{d_Y}[R_\beta(Y)]\|_\infty < \infty$.

It remains to verify global asymptotic stability of the limiting ODE $\dot{Q} = h_{\beta,\infty}(Q)$. Since $\mathcal{T}_0$ is a γ-contraction with unique fixed point 0, the argument of Lemma B.2 gives, for every $0 < \eta \leq L_\beta^{-1}$,

$$\|Q + \eta h_{\beta,\infty}(Q)\|_\infty \leq (1 - \eta\kappa_\beta)\|Q\|_\infty.$$

Let $Q(\cdot)$ solve $\dot{Q} = h_{\beta,\infty}(Q)$ and set $W(Q) := \|Q\|_\infty$. Since $h_{\beta,\infty}$ is globally Lipschitz,

$$Q(t + h) = Q(t) + h h_{\beta,\infty}(Q(t)) + r_t(h), \qquad \frac{\|r_t(h)\|_\infty}{h} \to 0 \quad (h \downarrow 0).$$

Using the 1-Lipschitz property of the sup norm, for sufficiently small $h > 0$,

$$W(Q(t + h)) \leq \|Q(t) + h h_{\beta,\infty}(Q(t))\|_\infty + \|r_t(h)\|_\infty \leq (1 - h\kappa_\beta) W(Q(t)) + \|r_t(h)\|_\infty.$$

Therefore

$$D^+W(Q(t)) := \limsup_{h\downarrow 0} \frac{W(Q(t+h)) - W(Q(t))}{h}$$
$$\leq -\kappa_\beta W(Q(t)) + \limsup_{h\downarrow 0} \frac{\|r_t(h)\|_\infty}{h} = -\kappa_\beta W(Q(t)). \tag{14}$$

Applying the comparison lemma (Khalil, 2002, Lemma 3.4) to the scalar ODE $\dot{u} = -\kappa_\beta u$ with $u(0) = W(x(0))$ yields

$$W(x(t)) \leq e^{-\kappa_\beta t} W(x(0)).$$

Since $W(x) = \|x\|_\infty$,

$$\|Q(t)\|_\infty \leq e^{-\kappa_\beta t} \|Q(0)\|_\infty,$$

so the origin is globally exponentially stable. This verifies Assumption 5.

Assumption 6: Markovian averaging. Assumption 6 of Liu et al. (2025) requires, for the relevant functions g and every initial state of the Markov chain,

$$\alpha_n \sum_{t=1}^{n} (g(Y_t) - \mathbb{E}_{d_Y}[g(Y)]) \to 0 \qquad \text{a.s.}$$

For $g = L_b$, this is immediate since $L_b \equiv 0$, and for $g = L$ it is immediate since L is constant. For fixed Q, the function $y \mapsto H_\beta(Q, y)$ is bounded on the finite state space. Since $\{Y_t\}$ is irreducible and aperiodic, Liu et al. (2025, Remark 2) implies Assumption 6 holds.

Global exponential stability of the limiting ODE. The mean drift $h_\beta(Q) = D(B+\beta I)(\mathcal{T}Q - Q)$ is globally Lipschitz since

$$\|h_\beta(Q) - h_\beta(Q')\|_\infty \leq (1+\gamma)\|D(B+\beta I)\|_\infty \|Q - Q'\|_\infty.$$

Hence $\dot{Q} = h_\beta(Q)$ has a unique global solution.

Set $x(t) := Q(t) - Q^\star$ and $W(x) := \|x\|_\infty$. Since $h_\beta(Q^\star) = 0$, Lemma B.2 gives, for every $0 < \eta \leq L_\beta^{-1}$,

$$\|x + \eta h_\beta(Q^\star + x)\|_\infty \leq (1 - \eta\kappa_\beta)\|x\|_\infty.$$

Along an ODE solution,

$$x(t+h) = x(t) + h h_\beta(Q^\star + x(t)) + r_t(h), \qquad \frac{\|r_t(h)\|_\infty}{h} \to 0.$$

Therefore, using again the 1-Lipschitz property of the sup norm,

$$W(x(t+h)) \leq (1 - h\kappa_\beta) W(x(t)) + \|r_t(h)\|_\infty$$

for all sufficiently small $h > 0$. Hence

$$D^+W(x(t)) := \limsup_{h\downarrow 0} \frac{W(x(t+h)) - W(x(t))}{h}$$
$$\leq -\kappa_\beta W(x(t)). \tag{15}$$

Applying the comparison lemma (Khalil, 2002, Lemma 3.4) to the scalar ODE $\dot{u} = -\kappa_\beta u$ with $u(0) = W(x(0))$ yields

$$W(x(t)) \leq e^{-\kappa_\beta t} W(x(0)).$$

Since $V(x) = \|x\|_\infty$,

$$\|Q(t) - Q^\star\|_\infty \leq e^{-\kappa_\beta t} \|Q(0) - Q^\star\|_\infty. \tag{16}$$

Hence $Q^\star$ is globally exponentially stable for the limiting ODE.

Theorem B.3 (Almost-sure convergence of V-SHARE under Markovian sampling)**.** *Suppose Assumption 1 holds, rewards are bounded, $\kappa_\beta > 0$, and*

$$\alpha_t = \frac{\alpha_0}{(t+h)^p}, \qquad \frac{1}{2} < p \leq 1.$$

Then

$$\sup_t \|Q_t\|_\infty < \infty \quad \text{a.s.}, \qquad Q_t \to Q^\star \quad \text{a.s.}$$

Proof. The preceding arguments verify Assumptions 1-6 of Liu et al. (2025). Their Theorem 7 therefore gives $\sup_t \|Q_t\|_\infty < \infty$ almost surely.

Their Corollary 8 implies that the asymptotic limit set of $\{Q_t\}$ is almost surely a bounded invariant set of the limiting ODE $\dot{Q} = h_\beta(Q)$. By equation 16, every trajectory of this ODE converges exponentially to $Q^\star$, so its global attractor is the singleton $\{Q^\star\}$. Hence the stochastic-approximation limit set is contained in $\{Q^\star\}$, and therefore $Q_t \to Q^\star$ almost surely. □

Remark 3. *The iterates generated by asynchronous Q-learning can be shown to be bounded a.s. by a simple induction argument (Gosavi, 2006). However, the same induction based argument doesn't work for* V-SHARE. *This is because the TD information pertaining to the state-action pair generated by the behvaiour policy is used to update the Q-values of multiple state-action pairs in* V-SHARE. *Hence, we require the ODE scaling arguments for the stability of iterates (Abounadi et al., 2002; Liu et al., 2025)*

B.4 Finite-time analysis under Markovian sampling

We now prove the finite-time guarantees for V-SHARE under the Markovian trajectory of Assumption 1. The argument follows the contractive Markovian stochastic-approximation framework of Chen et al. (2024). We first cast V-SHARE into their fixed-point form, then construct the generalized Moreau-envelope Lyapunov function, and finally specialize the resulting general recursion to different stepsize schedules.

Markovian contractive-SA representation. Recall the sample update map $H_\beta(Q, Y_{k+1})$ from Appendix B.3, so that

$$Q_{k+1} = Q_k + \alpha_k H_\beta(Q_k, Y_{k+1}),$$

and $\mathbb{E}_{d_Y}[H_\beta(Q, Y)] = h_\beta(Q)$. For an auxiliary analysis parameter $\eta > 0$, recall the V-SHARE fixed point map,

$$\mathsf{F}_{\eta,\beta}(Q, y) := Q + \eta H_\beta(Q, y), \qquad F_{\eta,\beta}(Q) := \mathbb{E}_{d_Y}[\mathsf{F}_{\eta,\beta}(Q, Y)] = Q + \eta h_\beta(Q).$$

Writing

$$\epsilon_k := \frac{\alpha_k}{\eta},$$

the V-SHARE recursion becomes

$$Q_{k+1} = Q_k + \epsilon_k \big[\mathsf{F}_{\eta,\beta}(Q_k, Y_{k+1}) - Q_k\big]. \tag{17}$$

By Lemma B.2, $F_{\eta,\beta}$ is a sup-norm contraction whenever $\kappa_\beta > 0$ and $0 < \eta \le L_\beta^{-1}$, with contraction factor

$$\rho_{\eta,\beta} := 1 - \eta\kappa_\beta.$$

We next verify the growth constants required by Chen et al. (2024). Let

$$C_\beta := 1 + \frac{\beta}{b_{\min}}.$$

From the sample update,

$$\|H_\beta(Q, y) - H_\beta(Q', y)\|_\infty \le C_\beta(1+\gamma)\|Q - Q'\|_\infty,$$

and therefore

$$\|\mathsf{F}_{\eta,\beta}(Q, y) - \mathsf{F}_{\eta,\beta}(Q', y)\|_\infty \le A_1\|Q - Q'\|_\infty, \qquad A_1 := 1 + \eta C_\beta(1+\gamma).$$

Moreover,

$$\|\mathsf{F}_{\eta,\beta}(0, y)\|_\infty \le B_1, \qquad B_1 := \eta C_\beta.$$

There is no additional martingale-difference term in equation 17, so in the notation of Chen et al. (2024) we may take $A_2 = B_2 = 0$. Hence define

$$A := A_1 + 1 = 2 + \eta C_\beta(1+\gamma), \qquad B := B_1 = \eta C_\beta,$$

and

$$c_1 := \left(\|Q_0 - Q^\star\|_\infty + \|Q_0\|_\infty + \frac{B}{A}\right)^2, \qquad c_2 := (A\|Q^\star\|_\infty + B)^2. \tag{18}$$

Mixing time. Let $\mathcal{M}_Y$ denote the finite-state Markov chain $\{Y_k\}$ and define its mixing time at precision $\delta > 0$ by

$$t_\delta := \min\left\{ m \geq 0 : \max_y \|P^m(y,\cdot) - d_Y\|_{\mathrm{TV}} \leq \delta \right\}.$$

For the effective stepsizes ϵ_k, write $t_k := t_{\epsilon_k}$ and, for $i \leq j$, $\epsilon_{i,j} := \sum_{\ell=i}^{j} \epsilon_\ell$. Under Assumption 1, $\mathcal{M}_Y$ is finite, irreducible and aperiodic, and hence geometrically mixing. In particular, $t_\delta = O(\log(1/\delta))$.

Generalized Moreau envelope. To exploit the sup-norm contraction of $F_{\eta,\beta}$, define $f(x) := \frac{1}{2}\|x\|_\infty^2$ and $g(x) := \frac{1}{2}\|x\|_p^2$, with $p = 2\log(|\mathcal{S}||\mathcal{A}|)$ and $d := |\mathcal{S}||\mathcal{A}|$. For $d \geq 3$, the ℓ_p norm satisfies

$$\ell_{cs}\|x\|_p \leq \|x\|_\infty \leq u_{cs}\|x\|_p, \qquad \ell_{cs} = \frac{1}{\sqrt{e}}, \qquad u_{cs} = 1,$$

and g is L-smooth with $L = p - 1 \leq 2\log(|\mathcal{S}||\mathcal{A}|)$.

For $\theta > 0$, define the generalized Moreau envelope

$$\mathcal{M}_\theta(x) := \min_{u \in \mathbb{R}^d} \left\{ \frac{1}{2}\|u\|_\infty^2 + \frac{1}{2\theta}\|x - u\|_p^2 \right\}.$$

The properties of the generalized Moreau envelope are summarized in Lemma E.1 and Corollary E.2.

Let $\rho := \rho_{\eta,\beta}$ and choose

$$\theta := \left(\frac{1+\rho}{2\rho}\right)^2 - 1.$$

The auxiliary constants of the Moreau analysis are

$$\phi_1 := \frac{1 + \theta u_{cs}^2}{1 + \theta \ell_{cs}^2}, \qquad \phi_2 := 1 - \rho\sqrt{\phi_1}, \qquad \phi_3 := \frac{114L(1 + \theta u_{cs}^2)}{\theta \ell_{cs}^2}. \tag{19}$$

For the sup norm, Corollary E.2 gives

$$\phi_1 \leq 3, \qquad \phi_2 \geq \frac{1-\rho}{2} = \frac{\eta\kappa_\beta}{2}, \qquad \phi_3 \leq \frac{456e\log(|\mathcal{S}||\mathcal{A}|)}{1-\rho} = \frac{456e\log(|\mathcal{S}||\mathcal{A}|)}{\eta\kappa_\beta}. \tag{20}$$

One-step Markovian Lyapunov recursion. The Markovian analysis is difficult because $\mathsf{F}_{\eta,\beta}(Q_k, Y_{k+1})$ is not conditionally unbiased given Q_k. Its effect is controlled by allowing the Markov chain to mix over the preceding t_k iterations.

Assume the effective stepsizes are nonincreasing and satisfy

$$\epsilon_{k-t_k,k-1} \leq \min\left\{ \frac{\phi_2}{\phi_3 A^2}, \frac{1}{4A} \right\}, \qquad k \geq t_k. \tag{21}$$

Then Lemma A.6 of Chen et al. (2024), specialized to equation 17, gives $\forall k > t_k$

$$\mathbb{E}[\mathcal{M}_\theta(Q_{k+1} - Q^\star)] \leq (1 - \phi_2\epsilon_k)\mathbb{E}[\mathcal{M}_\theta(Q_k - Q^\star)] + \frac{\phi_3 c_2}{2u_{cm}^2}\epsilon_k\epsilon_{k-t_k,k-1}.$$

Thus the first term gives contractive drift, while the second captures the additional error due to temporal dependence.

Let

$$K := \min\{k \geq 0 : k \geq t_k\}.$$

Using the norm equivalence of the Moreau envelope and repeatedly applying the preceding recursion gives, for every $k \geq K$,

$$\mathbb{E}\|Q_k - Q^\star\|_\infty^2 \leq \phi_1 c_1 \prod_{j=K}^{k-1}(1 - \phi_2\epsilon_j) + \phi_3 c_2 \sum_{i=K}^{k-1} \epsilon_i\epsilon_{i-t_i,i-1} \prod_{j=i+1}^{k-1}(1 - \phi_2\epsilon_j). \tag{22}$$

Equation 22 is the finite-time bound used below.

Theorem B.4 (Finite-time guarantees for V-SHARE under Markovian sampling)**.** *Suppose Assumption 1 holds and $\kappa_\beta > 0$. Set $\eta = L_\beta^{-1}$ and define the effective stepsize $\epsilon_k := \alpha_k L_\beta$. Define*

$$A_\beta := 2 + \frac{(1+\beta/b_{\min})(1+\gamma)}{L_\beta}, \qquad B_\beta := \frac{(1+\beta/b_{\min})}{L_\beta},$$

and

$$c_{1,\beta} := \left(\|Q_0 - Q^\star\|_\infty + \|Q_0\|_\infty + \frac{B_\beta}{A_\beta} \right)^2, \qquad c_{2,\beta} := \left(A_\beta \|Q^\star\|_\infty + B_\beta \right)^2.$$

Assume throughout that the effective stepsizes are nonincreasing and satisfy $\epsilon_{k-t_k,k-1} \le \min\left\{ \frac{\kappa_\beta^2}{912 e L_\beta^2 A_\beta^2 \log(|\mathcal{S}||\mathcal{A}|)}, \frac{1}{4A_\beta} \right\}$ for every $k \ge t_k$, where $\epsilon_{i,j} := \sum_{\ell=i}^{j} \epsilon_\ell$. Then:

(i) Constant stepsize. *If $\alpha_k \equiv \alpha$ and*

$$\alpha L_\beta \, t_{\alpha L_\beta} \le \min\left\{ \frac{\kappa_\beta^2}{912 e L_\beta^2 A_\beta^2 \log(|\mathcal{S}||\mathcal{A}|)}, \frac{1}{4A_\beta} \right\},$$

then, for every $k \ge t_{\alpha L_\beta}$,

$$\mathbb{E}\|Q_k - Q^\star\|_\infty^2 \le 3c_{1,\beta} \left(1 - \frac{\alpha\kappa_\beta}{2}\right)^{k - t_{\alpha L_\beta}} + 912 e \, \frac{\alpha L_\beta^3 c_{2,\beta} \log(|\mathcal{S}||\mathcal{A}|)}{\kappa_\beta^2} \, t_{\alpha L_\beta}. \tag{23}$$

In particular, geometric mixing gives $t_{\alpha L_\beta} = O(\log(1/(\alpha L_\beta)))$.

(ii) Inverse-linear stepsize. *Let*

$$\alpha_k = \frac{4}{\kappa_\beta (k+h)},$$

with h sufficiently large that the above mixing-window condition holds. Then, for every $k \ge K$,

$$\mathbb{E}\|Q_k - Q^\star\|_\infty^2 \le 3c_{1,\beta} \left(\frac{K+h}{k+h} \right)^2 + 58368 e^2 \frac{L_\beta^3 c_{2,\beta} \log(|\mathcal{S}||\mathcal{A}|)}{\kappa_\beta^3} \frac{t_k}{k+h}, \tag{24}$$

where

$$t_k = t_{\frac{4L_\beta}{\kappa_\beta(k+h)}}.$$

Hence, under geometric mixing,

$$\mathbb{E}\|Q_k - Q^\star\|_\infty^2 = O\left(\frac{\log k}{k} \right).$$

Proof. Fix $\eta = L_\beta^{-1}$ throughout. The effective Markovian-SA recursion has stepsizes $\epsilon_k = \alpha_k L_\beta$ and averaged contraction factor

$$\rho = 1 - \frac{\kappa_\beta}{L_\beta}.$$

For the generalized Moreau envelope introduced above, let ϕ_1, ϕ_2, ϕ_3 be the constants of Chen et al. (2024). By equation 20,

$$\phi_1 \le 3, \qquad \phi_2 \ge \frac{\kappa_\beta}{2L_\beta}, \qquad \phi_3 \le \frac{456 e L_\beta \log(|\mathcal{S}||\mathcal{A}|)}{\kappa_\beta}. \tag{25}$$

With $\eta = L_\beta^{-1}$, the sample-operator growth constants c_1, c_2 given in equation 18 are

$$A_\beta = 2 + \frac{(1+\beta/b_{\min})(1+\gamma)}{L_\beta}, \quad B_\beta = B_1 + B_2 = \frac{(1+\beta/b_{\min})}{L_\beta}$$

and consequently

$$c_{1,\beta} = \left(\|Q_0 - Q^\star\|_\infty + \|Q_0\|_\infty + \frac{B_\beta}{A_\beta} \right)^2, \quad c_{2,\beta} = \left(A_\beta \|Q^\star\|_\infty + B_\beta \right)^2.$$

The stepsize requirement of Chen et al. (2024) is

$$\epsilon_{k-t_k,k-1} \le \min\left\{\frac{\phi_2}{\phi_3 A_\beta^2}, \frac{1}{4A_\beta}\right\}.$$

Using equation 25,

$$\frac{\phi_2}{\phi_3 A_\beta^2} \ge \frac{\kappa_\beta^2}{912eL_\beta^2 A_\beta^2 \log(|\mathcal{S}||\mathcal{A}|)},$$

so the sufficient condition $\epsilon_{k-t_k,k-1} \le \min\left\{\frac{\kappa_\beta^2}{912eL_\beta^2 A_\beta^2 \log(|\mathcal{S}||\mathcal{A}|)}, \frac{1}{4A_\beta}\right\}$ used in the theorem implies the desired condition.

Their general Markovian-SA bound therefore gives, for every $k \ge K$,

$$\mathbb{E}\|Q_k - Q^\star\|_\infty^2 \le \phi_1 c_{1,\beta} \prod_{j=K}^{k-1}(1-\phi_2\epsilon_j) + \phi_3 c_{2,\beta} \sum_{i=K}^{k-1} \epsilon_i \epsilon_{i-t_i,i-1} \prod_{j=i+1}^{k-1}(1-\phi_2\epsilon_j). \tag{26}$$

We evaluate this expression for the various schedules separately.

(i) Constant stepsize. Let $\alpha_k \equiv \alpha$, so $\epsilon_k = \epsilon := \alpha L_\beta$, $t_k = t_\epsilon = t_{\alpha L_\beta}$, and $K = t_\epsilon$. Moreover, $\epsilon_{i-t_i,i-1} = \epsilon t_\epsilon$. Equation 26 gives

$$\mathbb{E}\|Q_k - Q^\star\|_\infty^2 \le \phi_1 c_{1,\beta}(1-\phi_2\epsilon)^{k-t_\epsilon} + \phi_3 c_{2,\beta}\epsilon^2 t_\epsilon \sum_{i=t_\epsilon}^{k-1}(1-\phi_2\epsilon)^{k-i-1}.$$

Since the geometric sum is at most $(\phi_2\epsilon)^{-1}$,

$$\mathbb{E}\|Q_k - Q^\star\|_\infty^2 \le \phi_1 c_{1,\beta}(1-\phi_2\epsilon)^{k-t_\epsilon} + \frac{\phi_3}{\phi_2} c_{2,\beta}\epsilon t_\epsilon.$$

Using equation 25 and $\epsilon = \alpha L_\beta$,

$$1-\phi_2\epsilon \le 1 - \frac{\alpha\kappa_\beta}{2}, \qquad \frac{\phi_3}{\phi_2}\epsilon \le 912e\frac{\alpha L_\beta^3 \log(|\mathcal{S}||\mathcal{A}|)}{\kappa_\beta^2}.$$

Together with $\phi_1 \le 3$, this gives equation 23.

(ii) Inverse-linear stepsize. Now let

$$\alpha_k = \frac{4}{\kappa_\beta(k+h)}, \qquad \epsilon_k = \frac{a}{k+h}, \qquad a := \frac{4L_\beta}{\kappa_\beta}.$$

The inverse-linear specialization of Chen et al. (2024) gives

$$\mathbb{E}\|Q_k - Q^\star\|_\infty^2 \le \phi_1 c_{1,\beta}\left(\frac{K+h}{k+h}\right)^{\phi_2 a} + \frac{8ea^2\phi_3 c_{2,\beta}}{\phi_2 a - 1}\frac{t_k}{k+h}.$$

By equation 25,

$$\phi_2 a \ge \frac{\kappa_\beta}{2L_\beta}\frac{4L_\beta}{\kappa_\beta} = 2,$$

so

$$\left(\frac{K+h}{k+h}\right)^{\phi_2 a} \le \left(\frac{K+h}{k+h}\right)^2, \qquad \phi_2 a - 1 \ge 1.$$

Furthermore,

$$8ea^2\phi_3 \le 8e\frac{16L_\beta^2}{\kappa_\beta^2}\frac{456eL_\beta\log(|\mathcal{S}||\mathcal{A}|)}{\kappa_\beta} = 58368e^2\frac{L_\beta^3\log(|\mathcal{S}||\mathcal{A}|)}{\kappa_\beta^3}.$$

Together with $\phi_1 \le 3$, this proves equation 24.

□

B.5 Dependence on β and comparison with Q-learning

We first examine how the finite-time constants of V-SHARE depend on the residual-preconditioning parameter β. Recall that

$$L_\beta = \max_{s,a} d(s)(\beta + \pi_b(a|s)), \qquad \kappa_\beta = \min_{s,a} d(s)\big[(1-\gamma)\beta - (1+\gamma) + 2\pi_b(a|s)\big].$$

Hence, for large β,

$$L_\beta \sim d_{\max}\beta, \qquad \kappa_\beta \sim d_{\min}(1-\gamma)\beta. \tag{27}$$

The Markovian finite-time bounds additionally depend on

$$A_\beta = 2 + \frac{(1+\beta/b_{\min})(1+\gamma)}{L_\beta}, \qquad B_\beta = \frac{(1+\beta/b_{\min})}{L_\beta},$$

and

$$c_{2,\beta} = \left(A_\beta \|Q^\star\|_\infty + B_\beta\right)^2.$$

Using equation 27,

$$A_\beta \longrightarrow 2 + \frac{1+\gamma}{d_{\max}b_{\min}}, \qquad B_\beta \longrightarrow \frac{1}{d_{\max}b_{\min}},$$

and therefore

$$c_{2,\beta} \longrightarrow c_{2,\infty} := \left[\left(2 + \frac{1+\gamma}{d_{\max}b_{\min}}\right)\|Q^\star\|_\infty + \frac{1}{d_{\max}b_{\min}}\right]^2. \tag{28}$$

In particular, $c_{2,\infty} = O(b_{\min}^{-2})$ when the remaining MDP parameters are held fixed.

Constant stepsizes. The stochastic radius in Theorem B.4(i) is proportional to

$$\frac{\alpha L_\beta^3 c_{2,\beta}}{\kappa_\beta^2}\, t_{\alpha L_\beta}.$$

By equation 27,

$$\frac{L_\beta^3 c_{2,\beta}}{\kappa_\beta^2} \sim \frac{d_{\max}^3}{d_{\min}^2(1-\gamma)^2}\,\beta\, c_{2,\infty}.$$

Thus, for fixed raw stepsize α, increasing β improves the deterministic contraction rate but increases the constant-stepsize stochastic radius linearly in β, up to the mixing-time factor. This is the contraction-variance tradeoff motivating the decreasing-β schedules of Section 3.4.

Inverse-linear stepsizes. For $\alpha_k = 4/[\kappa_\beta(k+h)]$, the leading stochastic coefficient in Theorem B.4(ii) is proportional to

$$\frac{L_\beta^3 c_{2,\beta}}{\kappa_\beta^3}.$$

Unlike the constant-stepsize radius, this quantity saturates as $\beta \to \infty$:

$$\frac{L_\beta^3 c_{2,\beta}}{\kappa_\beta^3} \longrightarrow \frac{d_{\max}^3}{d_{\min}^3(1-\gamma)^3} c_{2,\infty}. \tag{29}$$

Thus sufficiently large residual preconditioning improves the transient contraction without worsening the leading β-dependence of the inverse-linear asymptotic coefficient.

Comparison with asynchronous Q-learning. First, under constant stepsizes, asynchronous Q-learning has contraction gap $\mu_{\min}(1-\gamma)$ with $\mu_{\min} := \min_{s,a} d(s)\pi_b(a|s)$, whereas for large β, $\kappa_\beta \sim d_{\min}(1-\gamma)\beta$. Thus V-SHARE removes the multiplicative rare-action factor $b_{\min}$ from the leading deterministic contraction gap and hence obtains faster convergence as seen in Figure 1a.

Second, for inverse-linear stepsizes, Chen et al. (2024, Theorem B.1(1)(c)) obtain a stochastic prefactor proportional to $(1-\beta_1)^{-3} = [\mu_{\min}(1-\gamma)]^{-3}$ for asynchronous Q-learning. For V-SHARE, the corresponding coefficient scales as $L_\beta^3 c_{2,\beta}/\kappa_\beta^3$, which saturates for large β, with $c_{2,\beta} = O(b_{\min}^{-2})$. Hence, under the aligned-minima condition $\mu_{\min} = d_{\min}b_{\min}$, the same contractive-SA framework yields a $b_{\min}^{-3}$ dependence for asynchronous Q-learning and $b_{\min}^{-2}$ for large-β V-SHARE and thus improves the dependence on the rare-action from $b_{\min}^{-3}$ to $b_{\min}^{-2}$ when using inverse linear stepsizes.

C Decreasing Residual-Preconditioning Schedules

We first analyze the two-stage schedule separately. Unlike a genuinely time-varying schedule, each phase is a fixed contractive Markovian SA, so we may use the phase-specific analysis scale $\eta_j = L_j^{-1}$ in each phase. This yields a terminal stochastic radius determined by the low-β phase rather than by the largest residual weight used during learning.

Throughout, let $d := |\mathcal{S}||\mathcal{A}|$ and, for $j \in \{h, \ell\}$, write $L_j := L_{\beta_j}$, $\kappa_j := \kappa_{\beta_j}$, $\epsilon_j := \alpha L_j$, and $t_j := t_{\epsilon_j}$. Let $A_j, B_j, c_{2,j}$ denote the fixed-β_j constants from Theorem 3.2, and define

$$\nu_j := 912e \, \frac{\alpha L_j^3 c_{2,j} \log d}{\kappa_j^2} \, t_j.$$

We assume throughout that the fixed-β_j Markovian mixing-window condition of Theorem 3.2 holds for both phases.

C.1 Two-stage schedule

Consider

$$\beta_k = \begin{cases} \beta_h, & 0 \le k < K, \\ \beta_\ell, & k \ge K, \end{cases} \qquad \beta_h \ge \beta_\ell, \qquad \kappa_\ell > 0.$$

We assume $K \ge t_h$ and $T - K \ge t_\ell$.

Theorem C.1 (Two-stage residual preconditioning)**.** *Under the conditions above,*

$$\mathbb{E}\|Q_T - Q^\star\|_\infty^2 = O\Bigg(c_{1,h} \exp\left\{ -\frac{\alpha}{2} \left[\kappa_h (K - t_h) + \kappa_\ell (T - K - t_\ell) \right] \right\} + \nu_h \exp\left\{ -\frac{\alpha \kappa_\ell}{2} (T - K - t_\ell) \right\} + \nu_\ell \Bigg), \tag{30}$$

where $c_{1,h}$ is the initialization constant of Theorem 3.2 for β_h.

Consequently, after a sufficiently long low-β tail,

$$\mathbb{E}\|Q_T - Q^\star\|_\infty^2 = O(\nu_\ell) = O\left(\frac{\alpha L_\ell^3 c_{2,\ell} \log d}{\kappa_\ell^2} \, t_{\alpha L_\ell} \right).$$

Thus the terminal stochastic radius is the same order as that of running constant-β_ℓ V-SHARE*, while the transient retains the stronger contraction accumulated during the initial high-β phase.*

Proof. We apply the fixed-β analysis separately in the two phases.

For the first phase, Theorem 3.2 gives

$$\mathbb{E}\|Q_K - Q^\star\|_\infty^2 \le 3 c_{1,h} \exp\left\{ -\frac{\alpha \kappa_h}{2} (K - t_h) \right\} + \nu_h.$$

We now switch to β_ℓ and use the phase-specific analysis scale $\eta_\ell = L_\ell^{-1}$, so the effective SA stepsize is $\epsilon_\ell = \alpha L_\ell$. The first t_ℓ iterations after the switch form a short mixing burn-in. By Chen et al. (2024, Lemma A.2), since $\epsilon_\ell t_\ell \le 1/(4A_\ell)$,

$$\|Q_{K+t_\ell} - Q_K\|_\infty \le 2\epsilon_\ell t_\ell \left(A_\ell \|Q_K\|_\infty + B_\ell \right).$$

Writing $e_k := Q_k - Q^\star$ and using $\|Q_K\|_\infty \le \|e_K\|_\infty + \|Q^\star\|_\infty$, we obtain

$$\|e_{K+t_\ell}\|_\infty \le (1 + 2A_\ell \epsilon_\ell t_\ell) \|e_K\|_\infty + 2\epsilon_\ell t_\ell (A_\ell \|Q^\star\|_\infty + B_\ell).$$

Since $A_\ell \epsilon_\ell t_\ell \le 1/4$,

$$\mathbb{E}\|e_{K+t_\ell}\|_\infty^2 \le \frac{9}{2} \mathbb{E}\|e_K\|_\infty^2 + 8\epsilon_\ell^2 t_\ell^2 c_{2,\ell}.$$

From time $K + t_\ell$ onward, every mixing window lies entirely inside the low-β phase. Hence the fixed-β_ℓ one-step Moreau recursion from Appendix B.4 applies with its own constants $\phi_{1,\ell}, \phi_{2,\ell}, \phi_{3,\ell}$. Unrolling it gives

$$\mathbb{E}\|e_T\|_\infty^2 \le \phi_{1,\ell} \exp\left\{-\frac{\alpha\kappa_\ell}{2}(T - K - t_\ell)\right\} \mathbb{E}\|e_{K+t_\ell}\|_\infty^2 + \nu_\ell.$$

Using $\phi_{1,\ell} \le 3$ and substituting the preceding two displays yields

$$\begin{aligned}\mathbb{E}\|e_T\|_\infty^2 = O\Bigg(& c_{1,h} \exp\left\{-\frac{\alpha}{2}\left[\kappa_h(K - t_h) + \kappa_\ell(T - K - t_\ell)\right]\right\} \\ & + \nu_h \exp\left\{-\frac{\alpha\kappa_\ell}{2}(T - K - t_\ell)\right\} \\ & + \epsilon_\ell^2 t_\ell^2 c_{2,\ell} \exp\left\{-\frac{\alpha\kappa_\ell}{2}(T - K - t_\ell)\right\} + \nu_\ell\Bigg).\end{aligned}$$

Finally, the fixed-β_ℓ mixing-window condition, $\epsilon_\ell t_\ell = O(1)$, together with $\kappa_\ell \le L_\ell$, implies

$$\epsilon_\ell^2 t_\ell^2 c_{2,\ell} = O\left(\frac{\alpha L_\ell^3 c_{2,\ell} \log d}{\kappa_\ell^2} t_\ell\right) = O(\nu_\ell).$$

Absorbing this short burn-in contribution into ν_ℓ proves equation 30. □

Remark 4. *The bound separates the two effects cleanly. The initialization error benefits from contraction in both phases; stochastic error generated during the high-β phase is subsequently attenuated at the low-β contraction rate; and the eventual stochastic radius depends only on $(L_\ell, \kappa_\ell, c_{2,\ell})$. This is the advantage of analyzing the two fixed phases separately rather than treating the entire schedule with a single conservative analysis scale.*

C.2 General decreasing schedules

We now consider a deterministic decreasing sequence $\beta_k \in [\beta_\ell, \beta_h]$ with $\kappa_{\beta_\ell} > 0$. Ideally, at each iteration one would use the local analysis scale $\eta_k = L_{\beta_k}^{-1}$. This would produce the local effective stepsize $\epsilon_k = \alpha L_{\beta_k}$ and would retain the phase-specific stochastic scale $L_{\beta_k}^3 c_{2,\beta_k}/\kappa_{\beta_k}^2$. However, this leads to a genuinely time-varying contractive SA. In particular, the contraction factor $\rho_k = 1 - \kappa_{\beta_k}/L_{\beta_k}$ changes with k.

The generalized Moreau envelope used by Chen et al. (2024) is not independent of this contraction factor: For the sup norm $\|\cdot\|_\infty$, its smoothing parameter θ is chosen as a function of ρ_k. Thus using $\eta_k = L_{\beta_k}^{-1}$ at each iteration would produce a time-varying Lyapunov function $\mathcal{M}_k$. The standard one-step argument would then contain an additional variation term $\mathcal{M}_{k+1}(Q_{k+1} - Q^\star) - \mathcal{M}_k(Q_{k+1} - Q^\star)$, for which (Chen et al., 2024) provides no control. A contraction-factor-independent smooth approximation of $\frac{1}{2}\|\cdot\|_\infty^2$ would remove this difficulty, but such a construction is not available in the framework we use here.

We therefore use the conservative common scale $\eta := L_h^{-1}$, where $L_h := L_{\beta_h}$. The effective SA stepsize is then fixed: $\epsilon := \alpha L_h$. For each k, define

$$\mathsf{F}_k(Q, y) := Q + \frac{1}{L_h} H_{\beta_k}(Q, y),$$

whose stationary mean operator is

$$\overline{\mathsf{F}}_k(Q) = Q + \frac{1}{L_h} D(B + \beta_k I)(\mathcal{T}Q - Q).$$

This operator is a sup-norm contraction with factor $\rho_k = 1 - \kappa_k/L_h$, where $\kappa_k := \kappa_{\beta_k}$. Since $\kappa_k \ge \kappa_\ell$, all iterations are simultaneously controlled by the single worst-case contraction $\rho_\ell := 1 - \kappa_\ell/L_h$. We may therefore construct one fixed generalized Moreau envelope of $\|\cdot\|_\infty^2$ using ρ_ℓ and use it throughout the schedule.

Let ϕ_1, ϕ_3 be the corresponding common Moreau constants and $\phi_{2,k} := 1 - \rho_k\sqrt{\phi_1}$. Then $\phi_1 \le 3$, $\phi_{2,k} \ge \kappa_k/(2L_h)$, and $\phi_3 \le 456 e L_h \log d/\kappa_\ell$.

Define

$$A(\beta) := 2 + \frac{(1+\beta/b_{\min})(1+\gamma)}{L_h}, \qquad B(\beta) := \frac{(1+\beta/b_{\min})R_{\max}}{L_h},$$

and $c_2(\beta) := (A(\beta)\|Q^\star\|_\infty + B(\beta))^2$. Let $A_h := A(\beta_h)$, and, because the schedule is decreasing, set $\widetilde{c}_{2,k} := c_2(\beta_{(k-t_\epsilon)_+})$, where $t_\epsilon := t_{\alpha L_h}$.

A direct adaptation of Lemmas A.2-A.6 of Chen et al. (2024), using the fixed Moreau envelope above and the uniform growth constants corresponding to β_h, gives

$$\mathbb{E}\mathcal{M}(Q_{k+1} - Q^\star) \le (1 - \phi_{2,k}\epsilon)\mathbb{E}\mathcal{M}(Q_k - Q^\star) + \frac{\phi_3}{2u_{cm}^2}\widetilde{c}_{2,k}\epsilon^2 t_\epsilon,$$

provided

$$\epsilon t_\epsilon \le \min\left\{\frac{\kappa_\ell^2}{912eL_h^2A_h^2\log d}, \frac{1}{4A_h}\right\}.$$

Unrolling gives

$$\mathbb{E}\|Q_T - Q^\star\|_\infty^2 \le 3c_1 \exp\left\{-\frac{\alpha}{2}\sum_{k=t_\epsilon}^{T-1}\kappa_k\right\} + 456e\,\frac{\alpha^2 L_h^3 \log d}{\kappa_\ell} t_\epsilon \sum_{i=t_\epsilon}^{T-1}\widetilde{c}_{2,i}\exp\left\{-\frac{\alpha}{2}\sum_{j=i+1}^{T-1}\kappa_j\right\}.$$

Theorem C.2 (General decreasing residual-preconditioning schedule)**.** *Let $\{\beta_k\}_{k\ge 0}$ be any deterministic non-increasing sequence satisfying $\beta_k \in [\beta_\ell, \beta_h]$ and $\kappa_{\beta_\ell} > 0$. Let $L_h := L_{\beta_h}$, $\epsilon := \alpha L_h$, and $t_\epsilon := t_{\alpha L_h}$. Suppose*

$$\epsilon t_\epsilon \le \min\left\{\frac{\kappa_\ell^2}{912eL_h^2A_h^2\log d}, \frac{1}{4A_h}\right\},$$

where $\kappa_\ell := \kappa_{\beta_\ell}$ and $A_h := A(\beta_h)$.

Define

$$\widetilde{c}_{2,k} := c_2\left(\beta_{(k-t_\epsilon)_+}\right),$$

and

$$c_1 := \left(\|Q_0 - Q^\star\|_\infty + \|Q_0\|_\infty + \frac{B_h}{A_h}\right)^2, \qquad B_h := B(\beta_h).$$

Then, adapting the Markovian contractive-SA argument of Chen et al. (2024) to the deterministic sequence of sample operators $\{\mathsf{F}_k\}$ gives

$$\mathbb{E}\|Q_T - Q^\star\|_\infty^2 \le 3c_1 \exp\left\{-\frac{\alpha}{2}\sum_{k=t_\epsilon}^{T-1}\kappa_k\right\} + 456e\,\frac{\alpha^2 L_h^3 \log d}{\kappa_\ell} t_\epsilon \sum_{i=t_\epsilon}^{T-1}\widetilde{c}_{2,i}\exp\left\{-\frac{\alpha}{2}\sum_{j=i+1}^{T-1}\kappa_j\right\},$$

where $\kappa_k := \kappa_{\beta_k}$.

Proof. Using the fixed analysis scale $\eta = L_h^{-1}$ and the common generalized Moreau envelope constructed from the worst contraction factor $\rho_\ell = 1 - \kappa_\ell/L_h$, the proof follows the one-step Markovian Lyapunov argument of Chen et al. (2024, Lemmas A.2–A.6), with the sample-map growth constants taken uniformly over the mixing window. This gives

$$\mathbb{E}\mathcal{M}(Q_{k+1} - Q^\star) \le (1 - \phi_{2,k}\epsilon)\mathbb{E}\mathcal{M}(Q_k - Q^\star) + \frac{\phi_3}{2u_{cm}^2}\widetilde{c}_{2,k}\epsilon^2 t_\epsilon.$$

Unrolling the recursion, using $1 - x \le e^{-x}$, $\phi_1 \le 3$, $\phi_{2,k}\epsilon \ge \alpha\kappa_k/2$, and $\phi_3 \le 456eL_h\log d/\kappa_\ell$, yields the stated bound. □

Remark 5 (Conservatism of the general-schedule bound)**.** *Unlike the two-stage result, the bound above does not recover the fixed-β_ℓ terminal radius with L_ℓ^3. This is not an algorithmic effect, but a consequence of using a single Lyapunov function for the entire time-varying schedule: the common scale $\eta = L_h^{-1}$ propagates L_h into the stochastic coefficient. Obtaining a bound with the local factor L_{β_k} at each iteration would require a finite-time theory for time-varying contractive Markovian SA, or equivalently a way to control the variation of the associated Moreau envelopes. We leave such an extension for future work.*

C.3 SMOOTHLY DECREASING SCHEDULES

For explicit cumulative-contraction bounds, define $\underline{\kappa}_\beta := d_{\min}\big[(1-\gamma)\beta-(1+\gamma)+2b_{\min}\big]$. Since κ_β is the pointwise minimum of finitely many affine functions, $0 < \underline{\kappa}_\beta \le \kappa_\beta$ throughout the contraction regime.

We also use that $c_2(\beta)$ is Lipschitz on $[\beta_\ell, \beta_h]$. Thus, for some $L_c < \infty$, $|c_2(\beta) - c_2(\beta')| \le L_c|\beta - \beta'|$. Write $\underline{\kappa}_h := \underline{\kappa}_{\beta_h}$, $\underline{\kappa}_\ell := \underline{\kappa}_{\beta_\ell}$, and $c_{2,\ell} := c_2(\beta_\ell)$. Recall that $\epsilon = \alpha L_h$ and $t_\epsilon = t_{\alpha L_h}$.

Corollary C.3 (Linear decay)**.** *Consider*

$$\beta_k = \beta_\ell + (\beta_h - \beta_\ell)\left(1 - \frac{k}{T-1}\right).$$

Under the conditions of Theorem C.2,

$$\begin{aligned}\mathbb{E}\|Q_T - Q^\star\|_\infty^2 \le 3c_1 \exp\left\{-\frac{\alpha T}{4}(\underline{\kappa}_h + \underline{\kappa}_\ell) + O(\alpha t_\epsilon)\right\} \\ + O\left(\frac{\alpha L_h^3 c_{2,\ell} \log d}{\kappa_\ell^2} t_\epsilon\right) \\ + O\left(\frac{L_h^3 L_c \log d}{\kappa_\ell^3}\frac{t_\epsilon}{T}\right) + O\left(\frac{\alpha L_h^3 L_c \log d}{\kappa_\ell^2}\frac{t_\epsilon^2}{T}\right).\end{aligned} \tag{31}$$

Proof. Affineness of $\underline{\kappa}_\beta$ gives $\sum_{k=0}^{T-1} \underline{\kappa}_{\beta_k} = T(\underline{\kappa}_h + \underline{\kappa}_\ell)/2$. Since Theorem C.2 starts the contraction sum at t_ϵ, this contributes an $O(\alpha t_\epsilon)$ correction in the exponent. For $r := T-1-i$, $|\beta_{(i-t_\epsilon)_+} - \beta_\ell| = O((r + t_\epsilon)/T)$, and hence $\widetilde{c}_{2,i} \le c_{2,\ell} + O(L_c(r + t_\epsilon)/T)$. Using $\sum_{j=i+1}^{T-1} \kappa_j \ge \kappa_\ell r$, together with $\sum_{r\ge 0} e^{-\alpha\kappa_\ell r/2} = O((\alpha\kappa_\ell)^{-1})$ and $\sum_{r\ge 0} r e^{-\alpha\kappa_\ell r/2} = O((\alpha^2\kappa_\ell^2)^{-1})$, gives the result. □

Corollary C.4 (Polynomial decay)**.** *For $p > 0$, consider*

$$\beta_k = \beta_\ell + (\beta_h - \beta_\ell)\left(1 - \frac{k}{T-1}\right)^p.$$

Under the conditions of Theorem C.2,

$$\begin{aligned}\mathbb{E}\|Q_T - Q^\star\|_\infty^2 \le 3c_1 \exp\left\{-\frac{\alpha T}{2}\left[\underline{\kappa}_\ell + \frac{\underline{\kappa}_h - \underline{\kappa}_\ell}{p+1}\right] + O(\alpha t_\epsilon)\right\} \\ + O\left(\frac{\alpha L_h^3 c_{2,\ell} \log d}{\kappa_\ell^2} t_\epsilon\right) \\ + O\left(\frac{L_h^3 L_c \log d}{\kappa_\ell^{p+2}}\frac{t_\epsilon}{\alpha^{p-1}T^p}\right) + O\left(\frac{\alpha L_h^3 L_c \log d}{\kappa_\ell^2}\frac{t_\epsilon^{p+1}}{T^p}\right).\end{aligned} \tag{32}$$

Proof. A Riemann-sum estimate gives $\sum_{k=0}^{T-1} \underline{\kappa}_{\beta_k} = T[\underline{\kappa}_\ell + (\underline{\kappa}_h - \underline{\kappa}_\ell)/(p+1)] + O(1)$. Truncating the contraction sum at t_ϵ adds $O(\alpha t_\epsilon)$ to the exponent. Moreover, for $r := T-1-i$, $|\beta_{(i-t_\epsilon)_+} - \beta_\ell| = O((r + t_\epsilon)^p/T^p)$. For fixed $p > 0$, $(r + t_\epsilon)^p \le C_p(r^p + t_\epsilon^p)$. Using Lipschitz continuity of $c_2(\beta)$, $\sum_{j=i+1}^{T-1} \kappa_j \ge \kappa_\ell r$, $\sum_{r\ge 0} r^p e^{-\alpha\kappa_\ell r/2} = O((\alpha\kappa_\ell)^{-(p+1)})$, and $\sum_{r\ge 0} e^{-\alpha\kappa_\ell r/2} = O((\alpha\kappa_\ell)^{-1})$ in Theorem C.2 gives the result. □

Corollary C.5 (Normalized exponential decay)**.** *For $\lambda > 0$, define*

$$g_\lambda(x) := \frac{e^{-\lambda x} - e^{-\lambda}}{1 - e^{-\lambda}}, \qquad \beta_k = \beta_\ell + (\beta_h - \beta_\ell) g_\lambda\left(\frac{k}{T-1}\right),$$

and $A_\lambda := \int_0^1 g_\lambda(x)\,dx = \lambda^{-1} - (e^\lambda - 1)^{-1}$. Then, under the conditions of Theorem C.2,

$$\mathbb{E}\|Q_T - Q^\star\|_\infty^2 \le 3c_1 \exp\left\{-\frac{\alpha T}{2}\left[\underline{\kappa}_\ell + (\underline{\kappa}_h - \underline{\kappa}_\ell)A_\lambda\right] + O(\alpha t_\epsilon)\right\}$$

$$+ O\left(\frac{\alpha L_h^3 c_{2,\ell} \log d}{\kappa_\ell^2} t_\epsilon\right)$$
$$+ O_\lambda\left(\frac{L_h^3 L_c \log d}{\kappa_\ell^3} \frac{t_\epsilon}{T}\right) + O_\lambda\left(\frac{\alpha L_h^3 L_c \log d}{\kappa_\ell^2} \frac{t_\epsilon^2}{T}\right). \tag{33}$$

Proof. A Riemann-sum approximation gives $\sum_{k=0}^{T-1} \underline{\kappa}_{\beta_k} = T[\underline{\kappa}_\ell + (\underline{\kappa}_h - \underline{\kappa}_\ell)A_\lambda] + O(1)$. Truncating the contraction sum at t_ϵ contributes $O(\alpha t_\epsilon)$ in the exponent. Near the terminal point, $g_\lambda(1-u) = \lambda u/(e^\lambda - 1) + O_\lambda(u^2)$, so for $r := T - 1 - i$, $|\beta_{(i-t_\epsilon)_+} - \beta_\ell| = O_\lambda((r + t_\epsilon)/T)$. The remainder is identical to the linear-decay argument. □

D TWO-TIMESCALE V-SHARE: DETAILED PROOFS

Algorithm 2 V-SHARE

1: Initialize V_0, A_0, state S_0, stepsizes $\{\alpha_t, \eta_t\}_{t=0}^\infty$, and choose $\beta > 0$.
2: **for** $t = 0, 1, \ldots$ **do**
3: Sample $a_t \sim \pi_b(\cdot \mid s_t)$.
4: Observe $R_{t+1} = r(s_t, a_t)$ and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$.
5: Compute
$$\delta_t = R_{t+1} + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t).$$
6: Update, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$,
$$V_{t+1}(s) = V_t(s) + \mathbf{1}\{s_t = s\}\alpha_t \delta_t, \quad A_{t+1}(s,a) = A_t(s,a) + \eta_t \frac{\beta\, \mathbf{1}\{s_t = s, a_t = a\}}{\pi_b(a \mid s_t)} \delta_t.$$
7: **end for**

This appendix proves the two-timescale results stated in Section 3.5. Recall that $(LV)(s,a) = V(s)$, $B = \mathrm{blkdiag}(\mathbf{1} b_s^\top : s \in \mathcal{S})$, and $D_{(s,a),(s,a)} = d(s)$. The matrix B averages a state-action vector according to the behavior policy within each state and lifts the resulting statewise quantity to all actions at that state.

Throughout, $\beta > 0$ is fixed and
$$\tau_t := \beta \eta_t$$
denotes the effective advantage stepsize.

D.1 STOCHASTIC-APPROXIMATION FORM AND FROZEN FAST DYNAMICS

The two-timescale V-SHARE updates are
$$V_{t+1}(s_t) = V_t(s_t) + \alpha_t \delta_t, \qquad A_{t+1}(s_t, a) = A_t(s_t, a) + \tau_t \frac{\mathbf{1}\{a_t = a\}}{\pi_b(a|s_t)} \delta_t,$$
where $Q_t = LV_t + A_t$ and
$$\delta_t = r_{s_t, a_t} + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t).$$
We consider the regime
$$\frac{\alpha_t}{\tau_t} \to 0,$$
so that the advantage recursion evolves on the fast timescale and the value recursion on the slow timescale.

Let $Y_{t+1} := (s_t, a_t, s_{t+1})$ denote the Markovian observation process. Under Assumption 1, $\{Y_t\}$ is a finite-state irreducible and aperiodic Markov chain with invariant distribution
$$d_Y(s, a, s') = d(s)\pi_b(a|s)P(s'|s, a).$$

For $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $y = (s, a, s')$, define the sample temporal-difference residual

$$\delta(Q, y) := r(s, a) + \gamma \max_{a'} Q(s', a') - Q(s, a).$$

The sample advantage field is

$$H_A(Q, y)(s_0, a_0) := \mathbf{1}\{s_0 = s\} \frac{\mathbf{1}\{a_0 = a\}}{\pi_b(a|s)} \delta(Q, y),$$

while the lifted value field is

$$H_V(Q, y)(s_0, a_0) := \mathbf{1}\{s_0 = s\} \delta(Q, y).$$

The updates can therefore be written as

$$A_{t+1} = A_t + \tau_t H_A(Q_t, Y_{t+1}), \quad LV_{t+1} = LV_t + \alpha_t H_V(Q_t, Y_{t+1}).$$

Let $R(Q) := \mathcal{T}Q - Q$. The mean updates are obtained by integrating with respect to d_Y. For the advantage branch,

$$\begin{aligned}
\bar{H}_A(Q)(s, a) &:= \mathbb{E}_{Y \sim d_Y}[H_A(Q, Y)(s, a)] \\
&= \sum_{s'} d(s) \pi_b(a|s) P(s'|s, a) \frac{1}{\pi_b(a|s)} \Big[r(s, a) + \gamma \max_{a'} Q(s', a') - Q(s, a) \Big] \\
&= d(s) R(Q)(s, a).
\end{aligned}$$

Hence $\bar{H}_A(Q) = DR(Q)$. Similarly, for the lifted value branch,

$$\begin{aligned}
\bar{H}_V(Q)(s, a) &:= \mathbb{E}_{Y \sim d_Y}[H_V(Q, Y)(s, a)] \\
&= d(s) \sum_u \pi_b(u|s) \left[r(s, u) + \gamma \sum_{s'} P(s'|s, u) \max_{a'} Q(s', a') - Q(s, u) \right].
\end{aligned}$$

Since $B_s = \mathbf{e} b_s^\top$, this is exactly

$$\bar{H}_V(Q) = DBR(Q).$$

Thus the stationary mean updates of the two branches are

$$\bar{H}_A(Q) = DR(Q), \qquad \bar{H}_V(Q) = DBR(Q).$$

We next identify the frozen fast dynamics. Fix an arbitrary value vector V and consider the limiting ODE of the advantage update:

$$\dot{A} = D\big[\mathcal{T}(LV + A) - (LV + A)\big].$$

Its unique equilibrium is $\lambda(V) := Q^\star - LV$. Indeed, under the affine change of variables $Q = LV + A$, the frozen fast limiting ODE becomes $\dot{Q} = D(\mathcal{T}Q - Q)$. Define

$$F(Q) := Q + D(\mathcal{T}Q - Q).$$

For every Q, Q' and every (s, a),

$$\begin{aligned}
|F(Q)(s, a) - F(Q')(s, a)| &\le (1 - d(s))|Q(s, a) - Q'(s, a)| + d(s)|\mathcal{T}Q(s, a) - \mathcal{T}Q'(s, a)| \\
&\le \big[1 - d(s)(1 - \gamma)\big] \|Q - Q'\|_\infty.
\end{aligned}$$

Therefore

$$\|F(Q) - F(Q')\|_\infty \le (1 - \kappa)\|Q - Q'\|_\infty, \qquad \kappa := d_{\min}(1 - \gamma) > 0. \tag{34}$$

Let $Q(\cdot)$ solve the frozen fast limiting ODE and set $V(Q) := \|Q - Q^\star\|_\infty$. The preceding contraction equation 34 implies by the same arguments as equation 14,

$$D^+ V(Q(t)) \le -\kappa V(Q(t)).$$

Applying the comparison lemma (Khalil, 2002, Lemma 3.4) gives

$$\|Q(t) - Q^\star\|_\infty \le e^{-\kappa t} \|Q(0) - Q^\star\|_\infty.$$

Hence $Q^\star$ is the unique globally exponentially stable equilibrium of the frozen fast Q-dynamics, and therefore $\lambda(V) = Q^\star - LV$ is the unique globally exponentially stable equilibrium of the frozen fast advantage limiting ODE for every fixed V.

D.2 ALMOST-SURE TRACKING UNDER TIMESCALE SEPARATION

We now prove Theorem 3.4 under Markovian sampling. A standard two-timescale argument (Borkar, 2022) would study (A_t, V_t) jointly: for a frozen value iterate V, the fast advantage recursion has equilibrium $\lambda(V) = Q^\star - LV$, and the slow iterate evolves along this equilibrium manifold. In our setting, however, the reduced slow drift vanishes identically, since $LV + \lambda(V) = Q^\star$.

$$\dot{V} = DB(T(LV + \lambda(V)) - (LV + \lambda(V))) = 0$$

Thus every V is an equilibrium of the reduced slow ODE, reflecting the non-identifiability of the value-advantage decomposition. Standard two-timescale results (Mahadevan et al., 2026) requiring convergence of the slow iterate to a unique equilibrium are therefore unnecessarily strong for our purpose.

The key idea is that the fast tracking error (i.e. the error between the fast iterate and it's equilibrium) is exactly the Q-error:

$$\Delta_t := A_t - \lambda(V_t) = A_t - (Q^\star - LV_t) = Q_t - Q^\star.$$

After absorbing the fixed $\beta > 0$ into the effective fast gain $\tau_t := \beta\eta_t$, the recursion for Δ_t can be written as a single-timescale Markovian stochastic approximation with a vanishing perturbation. This strategy of writing the two-timescale stochastic approximation as a single-timescale stochastic approximation with a vanishing perturbation for the faster timescale was used in the proof of Borkar (2022, Lemma 8.1), for the martingale difference noise setting. We first establish a general extension of the Markovian ODE method of Liu et al. (2025) to such recursions and then specialize it to V-SHARE. This extension of the ODE method (Liu et al., 2025) to vanishing perturbations is analogous to Borkar (2022, Extension 3, Section 2.2), which was for martingale difference noise.

We also note that if the noise was indeed Martingale-difference noise i.e. each sample was obtained directly from the stationary distribution of the Markov chain, then Borkar (2022, Lemma 8.1) would imply

$$A_t - \lambda(V_t) = Q_t - Q^* \to 0 \quad a.s.$$

since $\lambda(V)$ is the globally exponentially stable equilibrium for the fast limiting ODE. In fact, the proof of (Borkar, 2022, Lemma 8.1) relies on results for single-timescale SA with vanishing perturbations (Borkar, 2022, Extension 3, Section 2.2).

D.2.1 MARKOVIAN STOCHASTIC APPROXIMATION WITH VANISHING PERTURBATIONS

Consider the recursion

$$x_{n+1} = x_n + a_n \left[H(x_n, Y_{n+1}) + r_n G(x_n, Y_{n+1})\right], \tag{35}$$

where $\{Y_n\}$ is a Markov chain and $r_n \geq 0$ is deterministic with $r_n \to 0$. Define $h(x) := \mathbb{E}_{d_Y}[H(x, Y)]$ and, for $c \geq 1$, $H_c(x, y) := H(cx, y)/c$. Let H_∞ and $h_\infty(x) := \mathbb{E}_{d_Y}[H_\infty(x, Y)]$ be the scaling limits appearing in Assumptions 3-5 of Liu et al. (2025).

Proposition D.1 (Markovian SA with a vanishing drift perturbation)**.** *Suppose Assumptions 1-5 and either Assumption 6 or Assumption 6′ of Liu et al. (2025) hold for the unperturbed recursion $x_{n+1} = x_n + a_n H(x_n, Y_{n+1})$. In particular, assume that the ODE $\dot{x} = h_\infty(x)$ has the origin as its globally asymptotically stable equilibrium.*

Assume additionally that, uniformly in y,

$$\|G(x, y) - G(x', y)\| \leq L_G\|x - x'\|, \qquad \sup_y \|G(0, y)\| \leq G_0$$

for some finite L_G, G_0. If $r_n \to 0$, then the iterates of equation 35 are almost surely stable:

$$\sup_n \|x_n\| < \infty \qquad a.s.$$

Moreover, their asymptotic limit set is almost surely a bounded invariant set of the unperturbed limiting ODE $\dot{x} = h(x)$. Consequently, if $x^\star$ is the unique globally asymptotically stable equilibrium of $\dot{x} = h(x)$, then $x_n \to x^\star$ almost surely.

Proof. We work on an almost-sure sample path on which all conclusions of the Markovian averaging lemmas of Liu et al. (2025) hold. The proof follows their proof of Theorem 7, with the additional task of showing that the $r_n G$ term disappears both in the stability rescaling and, after stability, in the ordinary ODE interpolation. We list the steps to our proof as follows:

Step 1: ODE-time blocks and stability rescaling. Following Liu et al. (2025), define

$$t(0) := 0, \qquad t(n) := \sum_{i=0}^{n-1} a_i, \qquad m(t) := \max\{n : t(n) \le t\}.$$

Fix an arbitrary ODE-time horizon $T > 0$ and define block endpoints

$$T_0 := 0, \qquad T_{j+1} := t(m(T_j + T) + 1)\,.$$

Let $\bar{x}(t) := x_{m(t)}$ be the piecewise-constant interpolation and set

$$R_j := \max\{1, \|\bar{x}(T_j)\|\}, \qquad \widehat{x}_j(s) := \frac{\bar{x}(T_j + s)}{R_j}, \qquad 0 \le s < T_j^+ := T_{j+1} - T_j.$$

Thus

$$\|\widehat{x}_j(0)\| \le 1$$

for every block, irrespective of whether the original iterates are bounded.

For $c \ge 1$, define the rescaled perturbation

$$G_c(x, y) := \frac{G(cx, y)}{c}.$$

The assumptions on G imply the uniform linear-growth estimate

$$\|G_c(x, y)\| \le L_G\|x\| + G_0, \qquad c \ge 1. \tag{36}$$

Indeed,

$$\frac{\|G(cx, y)\|}{c} \le \frac{\|G(cx, y) - G(0, y)\|}{c} + \frac{\|G(0, y)\|}{c} \le L_G\|x\| + G_0.$$

For

$$n_j := m(T_j), \qquad n_j(s) := m(T_j + s),$$

the normalized recursion on block j is

$$\widehat{x}_j(s) = \widehat{x}_j(0) + \sum_{i=n_j}^{n_j(s)-1} a_i \left[H_{R_j}\big(\widehat{x}_j(t(i) - T_j), Y_{i+1}\big) + r_i G_{R_j}\big(\widehat{x}_j(t(i) - T_j), Y_{i+1}\big)\right].$$

Step 2: uniform boundedness and equicontinuity of the normalized blocks. Since $r_n \to 0$, there exists N_0 such that $r_n \le 1$ for all $n \ge N_0$. For such n, Assumption 4 of Liu et al. (2025) and equation 36 give

$$\|H_c(x, y) + r_n G_c(x, y)\| \le \|H_c(0, y)\| + G_0 + \big(L(y) + L_G\big)\|x\|.$$

Thus the perturbed scaled recursion has the same affine linear-growth structure as the scaled recursion in the stability proof of Liu et al. (2025), with $L(y)$ replaced by $L(y) + L_G$ and the origin contribution enlarged by the deterministic constant G_0.

The finite-block estimates used by Liu et al. (2025) remain valid after these replacements. In particular, their Lemmas 37–39 provide, on the fixed almost-sure sample path, finite uniform bounds for

$$\sum_i a_i H_c(0, Y_{i+1}) \qquad \text{and} \qquad \sum_i a_i L(Y_{i+1})$$

over ODE-time blocks of fixed length. Adding L_G does not alter any centered Markovian averaging condition, since

$$[L(Y_i) + L_G] - [\mathbb{E}L(Y) + L_G] = L(Y_i) - \mathbb{E}L(Y),$$

and the additional origin contribution satisfies

$$G_0 \sum_i a_i = O_T(1)$$

on every fixed ODE-time block.

Consequently, the same discrete-Grönwall argument as in Liu et al. (2025, Lemma 40) yields, for every fixed $T > 0$, a finite sample-path-dependent constant $C_T(\omega)$ and an index $J_T(\omega)$ such that

$$\sup_{j \geq J_T} \sup_{0 \leq s \leq T} \|\widehat{x}_j(s)\| \leq C_T(\omega). \tag{37}$$

The finitely many blocks $j < J_T$ can be absorbed by increasing $C_T(\omega)$, so equation 37 may be taken to hold for all j. For brevity we henceforth write C_T for this finite sample-path-dependent constant.

The additional perturbation also preserves the extended equicontinuity argument of Liu et al. (2025, Lemma 15). Indeed, for $0 \leq s_1 \leq s_2 \leq T$,

$$\begin{aligned} &\left\| \sum_{i=n_j(s_1)}^{n_j(s_2)-1} a_i r_i G_{R_j}\big(\widehat{x}_j(t(i) - T_j), Y_{i+1}\big) \right\| \\ &\qquad \leq C_{G,T} \sum_{i=n_j(s_1)}^{n_j(s_2)-1} a_i, \end{aligned}$$

where

$$C_{G,T} := L_G C_T + G_0.$$

The ODE-time increment on the right converges uniformly to $|s_2 - s_1|$, up to the vanishing mesh $\sup_{i \geq n_j} a_i$. Hence the perturbation contributes a modulus of continuity that vanishes with $|s_2 - s_1|$, and $\{\widehat{x}_j(\cdot)\}_j$ is equicontinuous in the extended sense.

We will also need the analogue of the one-block growth estimate used in the proof of Liu et al. (2025, Lemma 16). Repeating the argument of Liu et al. (2025, Lemma 42) with the same modified growth bound gives finite sample-path-dependent constants C_1, C_2 such that

$$R_{j+1} \leq C_1 R_j + C_2 \qquad \text{for all sufficiently large } j. \tag{38}$$

After modifying C_1, C_2 to absorb finitely many initial blocks, equation 38 may be assumed to hold for every j.

Step 3: the perturbation vanishes under the stability rescaling. Using equation 37, for every $0 \leq s \leq T$,

$$\begin{aligned} &\left\| \sum_{i=n_j}^{n_j(s)-1} a_i r_i G_{R_j}\big(\widehat{x}_j(t(i) - T_j), Y_{i+1}\big) \right\| \\ &\qquad \leq C_{G,T} \left(\sup_{i \geq n_j} r_i \right) \sum_{i=n_j}^{n_j(T)-1} a_i. \end{aligned}$$

By the definition of $m(\cdot)$,

$$\sum_{i=n_j}^{n_j(T)-1} a_i \leq T + \sup_{i \geq n_j} a_i = T + o(1).$$

Since $n_j \to \infty$ and $r_n \to 0$,

$$\sup_{i \geq n_j} r_i \to 0.$$

Hence

$$\sup_{0 \leq s \leq T} \left\| \sum_{i=n_j}^{n_j(s)-1} a_i r_i G_{R_j}\big(\widehat{x}_j(t(i) - T_j), Y_{i+1}\big) \right\| \longrightarrow 0. \tag{39}$$

Thus the perturbation contributes zero to every subsequential limit of the stability-rescaled interpolations.

Step 4: stability. Suppose, toward a contradiction, that

$$\sup_j R_j = \infty.$$

The block-growth estimate equation 38 allows the same subsequence construction as in the proof of Liu et al. (2025, Lemma 16). Hence there exists a subsequence $\{j_k\}$ such that

$$R_{j_k} \longrightarrow \infty, \qquad R_{j_k+1} > R_{j_k} \quad \text{for every } k. \tag{40}$$

By the uniform boundedness and extended equicontinuity established in Step 2, Arzelà–Ascoli yields a further subsequence, not relabeled, along which the normalized trajectories $\widehat{x}_{j_k}(\cdot)$ converge uniformly on $[0, T]$.

We next identify their limit. Since neither H nor the driving Markov chain has been modified, the averaging and scaling arguments for the H-part are exactly those of Liu et al. (2025, Lemmas 17–19). These lemmas identify the limiting contribution of the H-increments with the ODE at infinity

$$\dot{z} = h_\infty(z).$$

The only additional term in the present recursion is the perturbation, and by equation 39 its cumulative contribution is $o(1)$ uniformly on $[0, T]$. Therefore the Grönwall argument used in Liu et al. (2025, Lemma 20) continues to show that every normalized subsequential limit of the perturbed recursion solves the same ODE

$$\dot{z} = h_\infty(z).$$

By Assumption 5 of Liu et al. (2025), the origin is globally asymptotically stable for this ODE. With the analogues of Lemmas 16 and 20 established above, the contraction argument of Liu et al. (2025, Lemma 21) applies verbatim: choosing T as in that proof, a sufficiently large normalized block must satisfy

$$R_{j_k+1} \le R_{j_k}$$

for some k, contradicting equation 40. Hence

$$\sup_j R_j < \infty.$$

It remains to pass from bounded block endpoints to bounded individual iterates. The within-block estimate of Liu et al. (2025, Lemma 22) depends only on the same finite-block growth bounds used above. Replacing $L(y)$ by $L(y) + L_G$ and enlarging the origin contribution by G_0 therefore gives the same conclusion:

$$\sup_n \|x_n\| < \infty \qquad \text{a.s.}$$

This proves stability of the perturbed recursion.

Step 5: the perturbation vanishes on ordinary two-sided ODE-time windows. Now that stability has been proved, fix the sample path and let $K(\omega)$ be a compact set containing all iterates $\{x_n(\omega)\}$. By global Lipschitzness of G,

$$C_K := \sup_{x \in K(\omega),\, y} \|G(x, y)\| \le L_G \sup_{x \in K(\omega)} \|x\| + G_0 < \infty.$$

Let $n_j \to \infty$ be arbitrary and fix $T > 0$. For the forward window, uniformly for $0 \le s \le T$,

$$\begin{aligned}
&\left\| \sum_{i=n_j}^{m(t(n_j)+s)-1} a_i r_i G(x_i, Y_{i+1}) \right\| \\
&\qquad \le C_K \left(\sup_{i \ge n_j} r_i \right) \sum_{i=n_j}^{m(t(n_j)+T)-1} a_i \\
&\qquad \le C_K \left(\sup_{i \ge n_j} r_i \right) \left(T + \sup_{i \ge n_j} a_i \right) \longrightarrow 0.
\end{aligned}$$

For the backward window, define, for all sufficiently large j,

$$\underline{n}_j := m(t(n_j) - T).$$

Then $\underline{n}_j \to \infty$. Uniformly for $-T \le s \le 0$,

$$\begin{aligned}
&\left\| \sum_{i=m(t(n_j)+s)}^{n_j-1} a_i r_i G(x_i, Y_{i+1}) \right\| \\
&\qquad \le C_K \left(\sup_{i \ge \underline{n}_j} r_i \right) \sum_{i=\underline{n}_j}^{n_j-1} a_i \\
&\qquad \le C_K \left(\sup_{i \ge \underline{n}_j} r_i \right) \left(T + \sup_{i \ge \underline{n}_j} a_i \right) \longrightarrow 0.
\end{aligned}$$

Consequently, the perturbation has zero asymptotic rate of change uniformly on every fixed two-sided ODE-time window $[-T, T]$.

Step 6: identification of the limiting ODE and invariant set. We now repeat the stable-interpolation argument used in the proof of Liu et al. (2025, Corollary 8). That argument considers arbitrary two-sided windows $[-T, T]$, extracts uniformly convergent subsequences, and identifies their limits as complete trajectories of the limiting ODE.

The Markovian averaging arguments for the H-increments are unchanged, because neither H nor $\{Y_n\}$ has been modified. Step 5 shows that, on every such two-sided window, the cumulative contribution of the additional $r_n G$ term converges uniformly to zero. Therefore the analogues of the interpolation and discretization arguments in the proof of Liu et al. (2025, Corollary 8) yield exactly the same limiting equation:

$$\dot{x} = h(x).$$

Hence every subsequential limit trajectory of the stable interpolation is a complete trajectory of the unperturbed ODE $\dot{x} = h(x)$. It follows, exactly as in the proof of Liu et al. (2025, Corollary 8), that the asymptotic limit set of $\{x_n\}$ is almost surely a nonempty bounded invariant set of

$$\dot{x} = h(x).$$

If this ODE has a unique globally asymptotically stable equilibrium $x^\star$, then its only bounded invariant set is $\{x^\star\}$. Consequently,

$$x_n \longrightarrow x^\star \qquad \text{a.s.}$$

□

D.2.2 Specialization to two-timescale V-SHARE

We now apply Proposition D.1 to the two-timescale V-SHARE recursion. Recall

$$V_{t+1}(s_t) = V_t(s_t) + \alpha_t \delta_t, \quad A_{t+1}(s_t, a) = A_t(s_t, a) + \tau_t \frac{\mathbf{1}\{a_t = a\}}{\pi_b(a|s_t)} \delta_t, \qquad \tau_t := \beta \eta_t.$$

Here

$$Q_t = LV_t + A_t, \qquad \delta_t = r(s_t, a_t) + \gamma \max_{a'} Q_t(s_{t+1}, a') - Q_t(s_t, a_t).$$

Let

$$\Delta_t := Q_t - Q^\star = A_t - (Q^\star - LV_t).$$

For $y = (s, a, s')$, define

$$\delta_\Delta(\Delta, y) := r(s, a) + \gamma \max_{a'} \big[Q^\star(s', a') + \Delta(s', a') \big] - Q^\star(s, a) - \Delta(s, a).$$

The action-specific fast update is

$$H_A(\Delta, y)(s_0, a_0) := \mathbf{1}\{s_0 = s\} \frac{\mathbf{1}\{a_0 = a\}}{\pi_b(a|s)} \delta_\Delta(\Delta, y),$$

while the shared-value update is

$$H_V(\Delta, y)(s_0, a_0) := \mathbf{1}\{s_0 = s\}\delta_\Delta(\Delta, y).$$

The error recursion is therefore

$$\Delta_{t+1} = \Delta_t + \tau_t \left[H_A(\Delta_t, Y_{t+1}) + \frac{\alpha_t}{\tau_t} H_V(\Delta_t, Y_{t+1}) \right]. \tag{41}$$

Thus equation 41 has precisely the form equation 35, with

$$a_t = \tau_t, \qquad r_t = \frac{\alpha_t}{\tau_t}, \qquad H = H_A, \qquad G = H_V.$$

Verification of the Markovian SA assumptions. We verify the assumptions of Proposition D.1 for the dominant field H_A.

Under Assumption 1, $Y_{t+1} = (s_t, a_t, s_{t+1})$ is a finite-state irreducible and aperiodic Markov chain with invariant distribution

$$d_Y(s, a, s') = d(s)\pi_b(a|s)P(s'|s, a).$$

For every Δ, Δ' and y,

$$|\delta_\Delta(\Delta, y) - \delta_\Delta(\Delta', y)| \le (1 + \gamma)\|\Delta - \Delta'\|_\infty.$$

Hence, with $b_{\min} := \min_{s,a} \pi_b(a|s) > 0$,

$$\|H_A(\Delta, y) - H_A(\Delta', y)\|_\infty \le \frac{1 + \gamma}{b_{\min}}\|\Delta - \Delta'\|_\infty,$$

and

$$\|H_V(\Delta, y) - H_V(\Delta', y)\|_\infty \le (1 + \gamma)\|\Delta - \Delta'\|_\infty.$$

Thus both sample fields are globally Lipschitz with linear growth. Since the Markov state space is finite and rewards are bounded, all required integrability conditions hold.

Averaging H_A under d_Y gives

$$h_A(\Delta) = D\left[\mathcal{T}(Q^\star + \Delta) - (Q^\star + \Delta)\right]. \tag{42}$$

Indeed, importance weighting cancels the behavior probability:

$$d(s)\pi_b(a|s)\frac{1}{\pi_b(a|s)} = d(s).$$

Similarly,

$$h_V(\Delta) = DB\left[\mathcal{T}(Q^\star + \Delta) - (Q^\star + \Delta)\right].$$

Only h_A determines the limiting ODE because $\alpha_t/\tau_t \to 0$.

For $c \ge 1$, define $H_{A,c}(\Delta, y) := H_A(c\Delta, y)/c$. Since $Q^\star$ and the reward are $O(1)$ under this scaling, $H_{A,c}$ converges uniformly on compact sets to

$$H_{A,\infty}(\Delta, y)(s_0, a) = \mathbf{1}\{s_0 = s\}\frac{\mathbf{1}\{u = a\}}{\pi_b(a|s)}\left[\gamma \max_{a'} \Delta(s', a') - \Delta(s, u)\right].$$

Its stationary average is

$$h_{A,\infty}(\Delta) = D(\mathcal{T}_0\Delta - \Delta), \tag{43}$$

where

$$(\mathcal{T}_0\Delta)(s, a) = \gamma \sum_{s'} P(s'|s, a) \max_{a'} \Delta(s', a').$$

The finite-state LLN/LIL conditions required by Liu et al. (2025) hold exactly as in Appendix B.3.

Global exponential stability of the dominant limiting ODE. Let

$$\kappa_{\text{fast}} := d_{\min}(1-\gamma).$$

Since $\mathcal{T}$ is a γ-contraction, the diagonally preconditioned Bellman map satisfies, for all $\zeta > 0, \quad \zeta < d_{\max}^{-1}$,

$$\|\Delta + \zeta h_A(\Delta)\|_\infty \le (1-\zeta\kappa_{\text{fast}})\|\Delta\|_\infty.$$

The proof is identical to the strategy used to prove equation 8. Let $\Delta(\cdot)$ solve $\dot{\Delta} = h_A(\Delta)$ and set $W(\Delta) := \|\Delta\|_\infty$. Since h_A is globally Lipschitz,

$$\Delta(t+h) = \Delta(t) + hh_A(\Delta(t)) + r_t(h), \qquad \frac{\|r_t(h)\|_\infty}{h} \to 0.$$

Using the 1-Lipschitz property of the sup norm and the preceding contraction inequality,

$$W(\Delta(t+h)) \le (1-h\kappa_{\text{fast}})W(\Delta(t)) + \|r_t(h)\|_\infty.$$

Therefore

$$D^+W(\Delta(t)) \le -\kappa_{\text{fast}}W(\Delta(t)).$$

Applying the comparison lemma (Khalil, 2002, Lemma 3.4) to $\dot{u} = -\kappa_{\text{fast}}u$ gives

$$\|\Delta(t)\|_\infty \le e^{-\kappa_{\text{fast}}t}\|\Delta(0)\|_\infty.$$

Hence the origin is the unique globally exponentially stable equilibrium of the dominant limiting ODE.

The same argument applies to the limiting ODE $\dot{\Delta} = h_{A,\infty}(\Delta)$ because $\mathcal{T}_0$ is a γ-contraction with unique fixed point 0. Thus

$$D^+\|\Delta(t)\|_\infty \le -\kappa_{\text{fast}}\|\Delta(t)\|_\infty,$$

and the origin is globally exponentially stable for the ODE at infinity. All stability hypotheses of Proposition D.1 are therefore satisfied.

Theorem D.2 (Almost-sure convergence of two-timescale V-SHARE)**.** *Fix any $\beta > 0$ and let $\tau_t := \beta\eta_t$. Suppose Assumption 1 holds and*

$$\tau_t = \frac{\tau_0}{(t+h)^p}, \qquad \alpha_t = \frac{\alpha_0}{(t+h)^q}, \qquad \frac{1}{2} < p < q \le 1.$$

Then

$$A_t - \big(Q^\star - LV_t\big) \to 0 \qquad \textit{a.s.},$$

and consequently

$$Q_t = LV_t + A_t \to Q^\star \qquad \textit{a.s.}$$

Proof. The recursion equation 41 satisfies Proposition D.1 with $a_t = \tau_t$ and

$$r_t = \frac{\alpha_t}{\tau_t} = \frac{\alpha_0}{\tau_0}(t+h)^{p-q} \longrightarrow 0,$$

since $p < q$. Moreover, $\tau_t = \tau_0/(t+h)^p$ satisfies the Markovian stochastic-approximation stepsize conditions because $1/2 < p \le 1$.

The dominant limiting ODE

$$\dot{\Delta} = h_A(\Delta)$$

has the origin as its unique globally exponentially stable equilibrium, and the corresponding limiting ODE at infinity is also globally exponentially stable, as shown above. The sample maps satisfy the required global Lipschitz and integrability assumptions, while the finite-state ergodic behavior chain satisfies the required Markov averaging conditions.

Proposition D.1 therefore gives

$$\Delta_t \to 0 \qquad \text{a.s.}$$

Since

$$\Delta_t = A_t - (Q^\star - LV_t) = Q_t - Q^\star,$$

both claims follow. □

Remark 6 (Why no threshold on β is required)**.** *The dominant fast dynamics in equation 41 are governed by H_A after the fixed $\beta > 0$ is absorbed into the effective gain $\tau_t = \beta\eta_t$. Consequently, the limiting ODE has contraction scale $d_{\min}(1-\gamma)$ and contains no residual-preconditioning threshold. The shared-value dynamics enter only through the vanishing perturbation α_t/τ_t. Thus, unlike single-timescale* V-SHARE*, the two-timescale result holds for every fixed $\beta > 0$.*

Remark 7 (Non-identifiability of the decomposition)**.** *The theorem does not assert that V_t or A_t converges individually. What converges is the action-value iterate $Q_t = LV_t + A_t$, which is the quantity of interest.*

D.3 Finite-time Markovian stochastic approximation with a small perturbation

We require the following extension of the contractive Markovian stochastic-approximation framework of Chen et al. (2024). The key difference is the presence of a state-dependent Markovian perturbation whose magnitude is controlled by a deterministic sequence $\{r_k\}$.

Consider the recursion

$$x_{k+1} = x_k + a_k \left[F(x_k, Y_k) - x_k + r_k G(x_k, Y_k)\right], \tag{44}$$

where $\{Y_k\}$ is a finite-state Markov chain, $\{a_k\}$ is a non-increasing stepsize sequence, and

$$0 \le r_k \le \bar{r} \qquad \text{for all } k. \tag{45}$$

Let μ denote the stationary distribution of $\{Y_k\}$ and define

$$\bar{F}(x) := \mathbb{E}_{Y\sim\mu}[F(x,Y)], \qquad \bar{G}(x) := \mathbb{E}_{Y\sim\mu}[G(x,Y)].$$

Throughout this subsection, $\|\cdot\|_c$ denotes an arbitrary norm on $\mathbb{R}^d$.

Assumption 2 (Dominant contractive operator)**.** *There exist $A_F, B_F > 0$ such that, for every $x, x' \in \mathbb{R}^d$ and $y \in \mathcal{Y}$,*

$$\|F(x,y) - F(x',y)\|_c \le A_F\|x - x'\|_c, \qquad \|F(0,y)\|_c \le B_F.$$

Moreover, $\bar{F}$ is a ρ-contraction in $\|\cdot\|_c$, for some $\rho \in (0,1)$, with unique fixed point $x^\star$:

$$\|\bar{F}(x) - \bar{F}(x')\|_c \le \rho\|x - x'\|_c, \qquad \bar{F}(x^\star) = x^\star.$$

Assumption 3 (Perturbation operator)**.** *There exist $A_G, B_G, L_G > 0$ such that, for every $x, x' \in \mathbb{R}^d$ and $y \in \mathcal{Y}$,*

$$\|G(x,y) - G(x',y)\|_c \le A_G\|x - x'\|_c, \qquad \|G(0,y)\|_c \le B_G,$$

and

$$\|\bar{G}(x) - \bar{G}(x')\|_c \le L_G\|x - x'\|_c.$$

We additionally assume that the perturbation preserves the target:

$$\bar{G}(x^\star) = 0. \tag{46}$$

Assumption 4 (Geometric mixing)**.** *The Markov chain $\{Y_k\}$ admits the unique stationary distribution μ and there exist $C_{\text{mix}} > 0$ and $\sigma \in (0,1)$ such that*

$$\max_{y\in\mathcal{Y}} \|P^m(y,\cdot) - \mu\|_{\text{TV}} \le C_{\text{mix}}\sigma^m, \qquad m \ge 0.$$

For $\delta > 0$, let

$$t_\delta := \min\left\{m \ge 0 : \max_y \|P^m(y,\cdot) - \mu\|_{\text{TV}} \le \delta\right\},$$

and define

$$t_k := \max\{1, t_{a_k}\}, \qquad a_{i,j} := \sum_{\ell=i}^{j} a_\ell.$$

We use the following constants. First, the complete sample increment in equation 44 satisfies an affine growth bound with

$$A_0 := A_F + 1 + \bar{r}A_G, \qquad B_0 := B_F + \bar{r}B_G. \tag{47}$$

For the Markovian-noise estimates, define the slightly enlarged constants

$$\widehat{A} := A_0 + A_G + L_G, \qquad \widehat{B} := B_0 + B_G. \tag{48}$$

In particular, $A_0 \le \widehat{A}$, $A_G + L_G \le \widehat{A}$, $B_0 \le \widehat{B}$, and $B_G \le \widehat{B}$.

Fixed Moreau envelope. Importantly, the Lyapunov function is constructed using only the contraction factor ρ of the dominant operator $\bar{F}$ and is therefore independent of r_k.

Let

$$f(x) := \frac{1}{2}\|x\|_c^2, \qquad g(x) := \frac{1}{2}\|x\|_s^2,$$

where g is L-smooth with respect to $\|\cdot\|_s$. Let $\ell_{cs}, u_{cs} > 0$ satisfy

$$\ell_{cs}\|x\|_s \le \|x\|_c \le u_{cs}\|x\|_s.$$

Choose $\theta > 0$ such that

$$\rho^2 < \frac{1+\theta\ell_{cs}^2}{1+\theta u_{cs}^2}, \tag{49}$$

and define

$$M(x) := \min_{u\in\mathbb{R}^d}\left\{\frac{1}{2}\|u\|_c^2 + \frac{1}{2\theta}\|x-u\|_s^2\right\}.$$

By the generalized Moreau-envelope properties used in Chen et al. (2024), there exists a norm $\|\cdot\|_m$ such that

$$M(x) = \frac{1}{2}\|x\|_m^2.$$

Moreover, with

$$\ell_{cm} := \sqrt{1+\theta\ell_{cs}^2}, \qquad u_{cm} := \sqrt{1+\theta u_{cs}^2},$$

we have

$$\ell_{cm}\|x\|_m \le \|x\|_c \le u_{cm}\|x\|_m.$$

Following Chen et al. (2024), define

$$\phi_1 := \frac{1+\theta u_{cs}^2}{1+\theta\ell_{cs}^2}, \qquad \phi_2 := 1-\rho\sqrt{\phi_1}, \qquad \phi_3 := \frac{114L(1+\theta u_{cs}^2)}{\theta\ell_{cs}^2}. \tag{50}$$

By equation 49, $\phi_2 > 0$.

Finally, define the perturbation-drift constant

$$C_G := 2L_G\frac{u_{cm}}{\ell_{cm}}., \tag{51}$$

and

$$\widetilde{\phi}_3 := \frac{(114+112\bar{r})Lu_{cm}^2}{\theta\ell_{cs}^2}. \tag{52}$$

We now state the finite-time extension.

Theorem D.3 (Contractive Markovian SA with a small Markovian perturbation)**.** *Suppose Assumptions 2- 4 hold. Assume*

$$\bar{r} \le \frac{\phi_2}{2C_G}, \tag{53}$$

and that the non-increasing stepsizes satisfy, for every $k \ge t_k$,

$$a_{k-t_k,k-1} \le \min\left\{\frac{\phi_2}{2\widetilde{\phi}_3\widehat{A}^2}, \frac{1}{4A_0}\right\}. \tag{54}$$

Let

$$K := \min\{k \ge 0 : k \ge t_k\},$$

and define

$$c_1 := \left(\|x_0 - x^\star\|_c + \|x_0\|_c + \frac{B_0}{A_0}\right)^2, \qquad c_2 := \left(\widehat{A}\|x^\star\|_c + \widehat{B}\right)^2. \tag{55}$$

Then, for every $k \ge K$,

$$\mathbb{E}\|x_k - x^\star\|_c^2 \le \phi_1 c_1 \prod_{j=K}^{k-1}(1-\phi_2 a_j) + \widetilde{\phi}_3 c_2 \sum_{i=K}^{k-1} a_i a_{i-t_i,i-1} \prod_{j=i+1}^{k-1}(1-\phi_2 a_j). \tag{56}$$

Thus the perturbation preserves the finite-time order of the original contractive Markovian-SA bound whenever its magnitude is uniformly small enough to be absorbed into the dominant negative drift.

Proof. Write

$$e_k := x_k - x^\star.$$

By the L/θ-smoothness of M and equation 44,

$$\begin{aligned} M(e_{k+1}) \le M(e_k) &+ \underbrace{a_k\langle \nabla M(e_k), \bar{F}(x_k) - x_k\rangle}_{T_1} + \underbrace{a_k r_k \langle \nabla M(e_k), \bar{G}(x_k)\rangle}_{T_1'} \\ &+ \underbrace{a_k\langle \nabla M(e_k), F(x_k, Y_k) - \bar{F}(x_k)\rangle}_{T_3} + \underbrace{a_k r_k \langle \nabla M(e_k), G(x_k, Y_k) - \bar{G}(x_k)\rangle}_{T_3'} \\ &+ \underbrace{\frac{La_k^2}{2\theta}\|F(x_k, Y_k) - x_k + r_k G(x_k, Y_k)\|_s^2}_{T_4}. \end{aligned} \tag{57}$$

We now bound each of the terms as follows:

Lemma D.4 (Dominant Moreau drift)**.**

$$T_1 \le -2\phi_2 a_k M(e_k) \quad \forall k \ge 1.$$

Proof. This is a direct restatement of Chen et al. (2024, Lemma A.1), applied to $e_k = x_k - x^\star$ and multiplied by a_k □

Lemma D.5 (Stationary perturbation drift)**.** *Suppose Assumption 3 holds and $\bar{G}(x^\star) = 0$. Then*

$$T_1' \le C_G a_k r_k M(e_k), \qquad C_G := 2L_G \frac{u_{cm}}{\ell_{cm}}.$$

Consequently, if

$$\bar{r} \le \frac{\phi_2}{2C_G},$$

then

$$T_1 + T_1' \le -\frac{3\phi_2}{2} a_k M(e_k).$$

Proof. Since $M(e) = \frac{1}{2}\|e\|_m^2$, we may write

$$\nabla M(e_k) = \|e_k\|_m v_{e_k}, \qquad v_{e_k} \in \partial\|e_k\|_m.$$

By Lemma E.1, $\|v_{e_k}\|_{m,*} \le 1$. Hence, by Hölder's inequality,

$$\begin{aligned} \langle \nabla M(e_k), \bar{G}(x_k)\rangle &= \|e_k\|_m \langle v_{e_k}, \bar{G}(x_k)\rangle \\ &\le \|e_k\|_m \|\bar{G}(x_k)\|_m. \end{aligned}$$

Since $\bar{G}(x^\star) = 0$, the Lipschitz assumption and norm equivalence give

$$\|\bar{G}(x_k)\|_m \le \frac{1}{\ell_{cm}}\|\bar{G}(x_k) - \bar{G}(x^\star)\|_c \le L_G \frac{u_{cm}}{\ell_{cm}}\|e_k\|_m.$$

Therefore

$$T_1' \le L_G \frac{u_{cm}}{\ell_{cm}} a_k r_k \|e_k\|_m^2 = 2L_G \frac{u_{cm}}{\ell_{cm}} a_k r_k M(e_k),$$

which proves the first claim. Combining this with Lemma D.4 and $r_k \le \bar{r} \le \phi_2/(2C_G)$ yields

$$T_1 + T_1' \le -\left(2\phi_2 - C_G\bar{r}\right) a_k M(e_k) \le -\frac{3\phi_2}{2} a_k M(e_k).$$

□

Lemma D.6 (Short-window iterate movement)**.** *Suppose*

$$\|F(x_k, Y_k) - x_k + r_k G(x_k, Y_k)\|_c \le A_0 \|x_k\|_c + B_0.$$

If

$$a_{k-t_k,k-1} \le \frac{1}{4A_0},$$

then, for every $j \in [k - t_k, k]$,

$$\|x_j - x_{k-t_k}\|_c \le 2a_{k-t_k,k-1} \left(A_0 \|x_{k-t_k}\|_c + B_0\right),$$

and

$$\|x_j - x_{k-t_k}\|_c \le 4a_{k-t_k,k-1} \left(A_0 \|x_k\|_c + B_0\right).$$

Proof. This is the iterate-movement argument of Chen et al. (2024, Lemma A.2). Its proof uses only the pathwise affine growth bound on the sample increment and the condition $a_{k-t_k,k-1} \le 1/(4A_0)$. Replacing the sample increment there by $F(x_k, Y_k) - x_k + r_k G(x_k, Y_k)$ and taking $A = A_0$, $B = B_0$ therefore gives the two stated bounds. □

Lemma D.7 (Dominant Markovian-noise term)**.** *Suppose the assumptions of Lemma D.6 hold and let* $\widehat{A}, \widehat{B}, c_2$ *be as defined above. Then*

$$\mathbb{E}[T_3] \le \frac{112L\widehat{A}^2 u_{cm}^2}{\theta \ell_{cs}^2} a_k a_{k-t_k,k-1} \mathbb{E}[M(e_k)] + \frac{56L}{\theta \ell_{cs}^2} a_k a_{k-t_k,k-1} c_2.$$

Proof. This is the Markovian-noise estimate of Chen et al. (2024, Lemmas A.3-A.4). Their proof uses only the smoothness of M, the short-window movement bounds, the Lipschitz and affine-growth bounds for F, and the mixing estimate for $\{Y_k\}$.

Lemma D.6 provides the required movement bounds for the present recursion, while $A_F + 1 \le \widehat{A}$ and $B_F \le \widehat{B}$. Thus the same calculation applies with the constants enlarged to $\widehat{A}, \widehat{B}$, yielding the stated inequality. □

Lemma D.8 (Perturbation Markovian-noise term)**.** *Under Assumption 3 and the conditions of Lemma D.6,*

$$\mathbb{E}[T_3'] \le r_k \frac{112L\widehat{A}^2 u_{cm}^2}{\theta \ell_{cs}^2} a_k a_{k-t_k,k-1} \mathbb{E}[M(e_k)] + r_k \frac{56L}{\theta \ell_{cs}^2} a_k a_{k-t_k,k-1} c_2.$$

Proof. Write $t := t_k$ and $a_{k-t,k-1} := \sum_{j=k-t}^{k-1} a_j$. As in the Markovian-noise decomposition of Chen et al. (2024, Lemma A.3), insert the delayed iterate x_{k-t} and decompose

$$T_3' = T_{31}' + T_{32}' + T_{33}',$$

where

$$\begin{aligned}
T_{31}' &:= a_k r_k \left\langle \nabla M(e_k) - \nabla M(e_{k-t}), G(x_k, Y_k) - \bar{G}(x_k) \right\rangle, \\
T_{32}' &:= a_k r_k \left\langle \nabla M(e_{k-t}), G(x_k, Y_k) - G(x_{k-t}, Y_k) \right. \\
&\qquad \left. + \bar{G}(x_{k-t}) - \bar{G}(x_k) \right\rangle, \\
T_{33}' &:= a_k r_k \left\langle \nabla M(e_{k-t}), G(x_{k-t}, Y_k) - \bar{G}(x_{k-t}) \right\rangle.
\end{aligned}$$

For convenience, define

$$S_k := \widehat{A}\|e_k\|_c + \widehat{A}\|x^\star\|_c + \widehat{B}.$$

By the definitions of $\widehat{A}, \widehat{B}$ and Lemma D.6,

$$\|x_k - x_{k-t}\|_c \le 4a_{k-t,k-1} S_k.$$

Moreover, using the reverse movement estimate exactly as in Chen et al. (2024, Lemma A.3),

$$\|e_{k-t}\|_c \le 2\left(\|e_k\|_c + \|x^\star\|_c + \frac{\widehat{B}}{\widehat{A}}\right).$$

Since M is L/θ-smooth with respect to $\|\cdot\|_s$,

$$\begin{aligned}\|\nabla M(e_k) - \nabla M(e_{k-t})\|_{s,*} &\le \frac{L}{\theta}\|x_k - x_{k-t}\|_s \\ &\le \frac{4L}{\theta\ell_{cs}} a_{k-t,k-1} S_k, \end{aligned} \tag{58}$$

and, because $\nabla M(0) = 0$,

$$\begin{aligned}\|\nabla M(e_{k-t})\|_{s,*} &\le \frac{L}{\theta}\|e_{k-t}\|_s \\ &\le \frac{2L}{\theta\ell_{cs}} \left(\|e_k\|_c + \|x^\star\|_c + \frac{\widehat{B}}{\widehat{A}} \right). \end{aligned} \tag{59}$$

We now bound the three terms separately.

Bounding T'_{31}. Since $\bar{G}(x^\star) = 0$, the sample growth and stationary Lipschitz bounds imply

$$\begin{aligned}\|G(x_k, Y_k) - \bar{G}(x_k)\|_c &\le \|G(x_k, Y_k)\|_c + \|\bar{G}(x_k)\|_c \\ &\le A_G\|x_k\|_c + B_G + L_G\|e_k\|_c \\ &\le \widehat{A}\|e_k\|_c + \widehat{A}\|x^\star\|_c + \widehat{B} = S_k. \end{aligned}$$

Hence, by Hölder's inequality, norm equivalence, and equation 58,

$$\begin{aligned} T'_{31} &\le \frac{a_k r_k}{\ell_{cs}} \|\nabla M(e_k) - \nabla M(e_{k-t})\|_{s,*} \|G(x_k, Y_k) - \bar{G}(x_k)\|_c \\ &\le r_k \frac{4L}{\theta\ell_{cs}^2} a_k a_{k-t,k-1} S_k^2. \end{aligned}$$

Using $(a+b)^2 \le 2a^2 + 2b^2$,

$$S_k^2 \le 2\widehat{A}^2\|e_k\|_c^2 + 2\left(\widehat{A}\|x^\star\|_c + \widehat{B}\right)^2.$$

Since $\|e_k\|_c^2 \le 2u_{cm}^2 M(e_k)$ and $c_2 := (\widehat{A}\|x^\star\|_c + \widehat{B})^2$, we obtain

$$T'_{31} \le r_k \frac{16L\widehat{A}^2 u_{cm}^2}{\theta\ell_{cs}^2} a_k a_{k-t,k-1} M(e_k) + r_k \frac{8L}{\theta\ell_{cs}^2} a_k a_{k-t,k-1} c_2. \tag{60}$$

Bounding T'_{32}. By the sample and stationary Lipschitz properties,

$$\begin{aligned}\|G(x_k, Y_k) - G(x_{k-t}, Y_k) + \bar{G}(x_{k-t}) - \bar{G}(x_k)\|_c &\le (A_G + L_G)\|x_k - x_{k-t}\|_c \\ &\le 2\widehat{A}\|x_k - x_{k-t}\|_c \\ &\le 8\widehat{A} a_{k-t,k-1} S_k, \end{aligned}$$

where in the second inequality we used $A_G, L_G \le \widehat{A}$.

Combining this with Hölder's inequality, equation 59, and norm equivalence gives

$$\begin{aligned} T'_{32} &\le \frac{a_k r_k}{\ell_{cs}} \|\nabla M(e_{k-t})\|_{s,*} \times \|G(x_k, Y_k) - G(x_{k-t}, Y_k) + \bar{G}(x_{k-t}) - \bar{G}(x_k)\|_c \\ &\le r_k \frac{16L}{\theta\ell_{cs}^2} a_k a_{k-t,k-1} S_k^2. \end{aligned}$$

Applying the same bound on S_k^2 as above yields

$$T'_{32} \le r_k \frac{64L\widehat{A}^2 u_{cm}^2}{\theta\ell_{cs}^2} a_k a_{k-t,k-1} M(e_k) + r_k \frac{32L}{\theta\ell_{cs}^2} a_k a_{k-t,k-1} c_2. \tag{61}$$

Bounding T'_{33}. Let $\mathcal{F}_{k-t}$ denote the sigma-field generated by the history through time $k-t$. Since x_{k-t} and Y_{k-t} are $\mathcal{F}_{k-t}$-measurable, the Markov property gives

$$\mathbb{E}[G(x_{k-t}, Y_k) \,|\, \mathcal{F}_{k-t}] - \bar{G}(x_{k-t}) = \sum_{y \in \mathcal{Y}} \big(P^t(Y_{k-t}, y) - \mu(y)\big)\, G(x_{k-t}, y).$$

Therefore, using the definition of total-variation distance,

$$\big\|\mathbb{E}[G(x_{k-t}, Y_k) \mid \mathcal{F}_{k-t}] - \bar{G}(x_{k-t})\big\|_c \le 2\|P^t(Y_{k-t}, \cdot) - \mu\|_{\mathrm{TV}} \sup_{y \in \mathcal{Y}} \|G(x_{k-t}, y)\|_c.$$

By the definition $t = t_{a_k}$,

$$\|P^t(Y_{k-t}, \cdot) - \mu\|_{\mathrm{TV}} \le a_k.$$

Moreover,

$$\|G(x_{k-t}, y)\|_c \le A_G \|x_{k-t}\|_c + B_G.$$

Using the reverse movement estimate and the definitions of $\widehat{A}, \widehat{B}$ exactly as above,

$$A_G \|x_{k-t}\|_c + B_G \le 2S_k.$$

Consequently,

$$\big\|\mathbb{E}[G(x_{k-t}, Y_k) \mid \mathcal{F}_{k-t}] - \bar{G}(x_{k-t})\big\|_c \le 4a_k S_k. \tag{62}$$

Taking conditional expectation in T'_{33} and applying Hölder's inequality, equation 59, and equation 62,

$$\begin{aligned}\mathbb{E}[T'_{33} \mid \mathcal{F}_{k-t}] &\le \frac{a_k r_k}{\ell_{cs}} \|\nabla M(e_{k-t})\|_{s,*} \times \big\|\mathbb{E}[G(x_{k-t}, Y_k) \mid \mathcal{F}_{k-t}] - \bar{G}(x_{k-t})\big\|_c \\ &\le r_k \frac{8L}{\theta \ell_{cs}^2} a_k^2 S_k^2.\end{aligned}$$

Thus,

$$\mathbb{E}[T'_{33}] \le r_k \frac{32L\widehat{A}^2 u_{cm}^2}{\theta \ell_{cs}^2} a_k^2 \mathbb{E}[M(e_k)] + r_k \frac{16L}{\theta \ell_{cs}^2} a_k^2 c_2. \tag{63}$$

Finally, since $t_k \ge 1$ and $\{a_k\}$ is non-increasing,

$$a_k \le a_{k-t_k, k-1}.$$

Hence equation 63 can be bounded using $a_k^2 \le a_k a_{k-t_k,k-1}$. Summing equation 60, equation 61, and equation 63, and using

$$16 + 64 + 32 = 112, \qquad 8 + 32 + 16 = 56,$$

gives

$$\mathbb{E}[T'_3] \le r_k \frac{112L\widehat{A}^2 u_{cm}^2}{\theta \ell_{cs}^2} a_k a_{k-t_k,k-1} \mathbb{E}[M(e_k)] + r_k \frac{56L}{\theta \ell_{cs}^2} a_k a_{k-t_k,k-1} c_2,$$

as claimed. □

Lemma D.9 (Second-order Moreau remainder)**.** *Under the uniform increment bound*

$$\|F(x_k, Y_k) - x_k + r_k G(x_k, Y_k)\|_c \le A_0 \|x_k\|_c + B_0,$$

we have

$$T_4 \le \frac{2L\widehat{A}^2 u_{cm}^2}{\theta \ell_{cs}^2} a_k a_{k-t_k,k-1} M(e_k) + \frac{L}{\theta \ell_{cs}^2} a_k a_{k-t_k,k-1} c_2.$$

Proof. By norm equivalence,

$$\begin{aligned}T_4 &= \frac{La_k^2}{2\theta} \|F(x_k, Y_k) - x_k + r_k G(x_k, Y_k)\|_s^2 \\ &\le \frac{La_k^2}{2\theta \ell_{cs}^2} (A_0 \|x_k\|_c + B_0)^2.\end{aligned}$$

Using $(a+b)^2 \le 2a^2 + 2b^2$, $A_0 \le \widehat{A}$, the norm equivalence $\|e_k\|_c^2 \le 2u_{cm}^2 M(e_k)$, and the definition of c_2 gives

$$T_4 \le \frac{2L\widehat{A}^2 u_{cm}^2}{\theta \ell_{cs}^2} a_k^2 M(e_k) + \frac{L}{\theta \ell_{cs}^2} a_k^2 c_2.$$

Finally, $a_k \le a_{k-t_k,k-1}$ implies the stated result. □

One-step Lyapunov recursion. Combining Lemmas D.4, D.5, D.7, D.8, and D.9, and using $r_k \le \bar{r}$,, gives

$$\mathbb{E}[M(e_{k+1})] \le \left[1 - \frac{3\phi_2}{2}a_k + \frac{(114+112\bar{r})L\widehat{A}^2 u_{cm}^2}{\theta \ell_{cs}^2} a_k a_{k-t_k,k-1}\right]\mathbb{E}[M(e_k)] + \frac{(57+56\bar{r})L}{\theta\ell_{cs}^2} a_k a_{k-t_k,k-1} c_2. \tag{64}$$

By equation 52, this is

$$\mathbb{E}[M(e_{k+1})] \le \left[1 - \frac{3\phi_2}{2}a_k + \widetilde{\phi}_3 \widehat{A}^2 a_k a_{k-t_k,k-1}\right]\mathbb{E}[M(e_k)] + \frac{\widetilde{\phi}_3}{2u_{cm}^2} a_k a_{k-t_k,k-1} c_2. \tag{65}$$

Under equation 54,

$$\widetilde{\phi}_3 \widehat{A}^2 a_{k-t_k,k-1} \le \frac{\phi_2}{2},$$

and hence

$$\mathbb{E}[M(e_{k+1})] \le (1-\phi_2 a_k)\mathbb{E}[M(e_k)] + \frac{\widetilde{\phi}_3}{2u_{cm}^2} a_k a_{k-t_k,k-1} c_2. \tag{66}$$

Unrolling the recursion. Recursively applying equation 66 from K to $k-1$ gives

$$\mathbb{E}[M(e_k)] \le \mathbb{E}[M(e_K)] \prod_{j=K}^{k-1}(1-\phi_2 a_j) + \frac{\widetilde{\phi}_3 c_2}{2u_{cm}^2} \sum_{i=K}^{k-1} a_i a_{i-t_i,i-1} \prod_{j=i+1}^{k-1}(1-\phi_2 a_j).$$

Since

$$\|e_k\|_c^2 \le 2u_{cm}^2 M(e_k), \qquad M(e_K) \le \frac{1}{2\ell_{cm}^2}\|e_K\|_c^2,$$

we obtain

$$\mathbb{E}\|e_k\|_c^2 \le \phi_1 \mathbb{E}\|e_K\|_c^2 \prod_{j=K}^{k-1}(1-\phi_2 a_j) + \widetilde{\phi}_3 c_2 \sum_{i=K}^{k-1} a_i a_{i-t_i,i-1} \prod_{j=i+1}^{k-1}(1-\phi_2 a_j). \tag{67}$$

Finally, during the initial interval $k < K$, the affine growth bound with constants (A_0, B_0) and the proof of Chen et al. (2024, Lemma A.2) give

$$\|x_K - x^\star\|_c^2 \le \left(\|x_0 - x^\star\|_c + \|x_0\|_c + \frac{B_0}{A_0}\right)^2 = c_1.$$

Substituting this into equation 67 proves equation 56. □

Remark 8 (Interpretation). *Relative to the analysis of Chen et al. (2024), the perturbation introduces exactly two new first-order terms. The stationary component T_1' consumes part of the negative drift and is controlled by the uniform smallness condition on $\bar{r}$, while the centered Markovian component T_3' contributes another mixing-window error of the same order $a_k a_{k-t_k,k-1}$ as the original Markovian noise. Consequently, a sufficiently small state-dependent Markovian perturbation changes only the constants, and not the finite-time order of the contractive Markovian-SA bound.*

D.4 Finite-time tracking under timescale separation

We now apply Theorem D.3 to the two-timescale V-SHARE recursion. Recall the tracking error

$$\Delta_t := A_t - \lambda(V_t) = LV_t + A_t - Q^\star = Q_t - Q^\star.$$

Define

$$H(\Delta) := \mathcal{T}(Q^\star + \Delta) - (Q^\star + \Delta).$$

Since $\mathcal{T}Q^\star = Q^\star$ and $\mathcal{T}$ is a γ-contraction,

$$\|H(\Delta) - H(\Delta')\|_\infty \le (1+\gamma)\|\Delta - \Delta'\|_\infty, \qquad \|H(\Delta)\|_\infty \le (1+\gamma)\|\Delta\|_\infty.$$

For $y = (s, a, s')$, let

$$\delta_\Delta(\Delta, y) := r(s,a) + \gamma \max_{a'} \big[Q^\star(s', a') + \Delta(s', a')\big] - Q^\star(s,a) - \Delta(s,a).$$

Define the fast and slow sample operators by

$$H_A(\Delta, y)(s_0, a_0) := \mathbf{1}\{s_0 = s\} \frac{\mathbf{1}\{a_0 = a\}}{\pi_b(a|s)} \delta_\Delta(\Delta, y),$$

and

$$H_V(\Delta, y)(s_0, a_0) := \mathbf{1}\{s_0 = s\} \delta_\Delta(\Delta, y).$$

Writing $r_t := \alpha_t / \tau_t$, the tracking error satisfies the recursion

$$\Delta_{t+1} = \Delta_t + \tau_t \left[H_A(\Delta_t, Y_{t+1}) + r_t H_V(\Delta_t, Y_{t+1})\right]. \tag{68}$$

Reduction to perturbed Markovian SA. To match Theorem D.3, set

$$F(\Delta, y) := \Delta + H_A(\Delta, y), \qquad G(\Delta, y) := H_V(\Delta, y).$$

Under the invariant distribution of the Markov chain,

$$\bar{F}(\Delta) = \Delta + DH(\Delta), \qquad \bar{G}(\Delta) = DBH(\Delta).$$

Both operators vanish at the target $\Delta^\star = 0$.

Let

$$\kappa := d_{\min}(1 - \gamma).$$

The dominant stationary operator satisfies

$$\|\bar{F}(x) - \bar{F}(y)\|_\infty \le (1 - \kappa)\|x - y\|_\infty.$$

Hence the fixed contraction factor used throughout the Moreau analysis is

$$\rho := 1 - \kappa.$$

For the perturbation,

$$\|\bar{G}(x) - \bar{G}(y)\|_\infty \le L_G \|x - y\|_\infty, \qquad L_G := d_{\max}(1 + \gamma).$$

Sample regularity. Let

$$b_{\min} := \min_{s,a} \pi_b(a|s), \qquad G_\star := 1 + (1 + \gamma)\|Q^\star\|_\infty,$$

where we used $r(s,a) \in [0,1]$.

The fast sample operator satisfies

$$\|H_A(x, y) - H_A(x', y)\|_\infty \le \frac{1 + \gamma}{b_{\min}} \|x - x'\|_\infty,$$

and

$$\|H_A(0, y)\|_\infty \le \frac{G_\star}{b_{\min}}.$$

Likewise,

$$\|H_V(x, y) - H_V(x', y)\|_\infty \le (1 + \gamma)\|x - x'\|_\infty, \qquad \|H_V(0, y)\|_\infty \le G_\star.$$

Therefore, in the notation of Theorem D.3,

$$A_F = 1 + \frac{1 + \gamma}{b_{\min}}, \qquad B_F = \frac{G_\star}{b_{\min}},$$

and

$$A_G = 1 + \gamma, \qquad B_G = G_\star.$$

Suppose $r_t \le r_{\max}$ for all t. Define

$$A_0 := A_F + 1 + r_{\max} A_G, \qquad B_0 := B_F + r_{\max} B_G,$$

and

$$\widehat{A} := A_0 + A_G + L_G, \qquad \widehat{B} := B_0 + B_G.$$

Fixed Moreau envelope and timescale separation. Let

$$d := |\mathcal{S}||\mathcal{A}|, \qquad p := 2\log d, \qquad g(x) := \frac{1}{2}\|x\|_p^2.$$

We use the ℓ_∞ specialization of the generalized Moreau envelope given in Corollary E.2. In particular, with

$$\theta := \left(\frac{1+\rho}{2\rho}\right)^2 - 1,$$

we may choose

$$\ell_{cs} = e^{-1/2}, \qquad u_{cs} = 1,$$

and obtain

$$\phi_1 \le 3, \qquad \phi_2 \ge \frac{\kappa}{2}, \qquad \phi_3 \le \frac{456e\log d}{\kappa}.$$

Let

$$\ell_{cm} := \sqrt{1+\theta\ell_{cs}^2}, \qquad u_{cm} := \sqrt{1+\theta u_{cs}^2}.$$

By Lemma D.5, the stationary perturbation satisfies

$$T_1' \le C_G \tau_t r_t M(\Delta_t), \qquad C_G := 2L_G \frac{u_{cm}}{\ell_{cm}}.$$

Hence we impose the uniform timescale-separation condition

$$\frac{\alpha_t}{\tau_t} = r_t \le r_{\max} \le c_{\text{sep}} := \frac{\phi_2}{2C_G} = \frac{\phi_2 \ell_{cm}}{4L_G u_{cm}}, \qquad \forall t. \tag{69}$$

Under this condition, the slow value update consumes at most one quarter of the dominant negative Moreau drift.

The Markovian-noise and second-order terms are controlled by the constant

$$\widetilde{\phi}_3 := \frac{(114+112r_{\max})Lu_{cm}^2}{\theta\ell_{cs}^2}.$$

Using Corollary E.2,

$$\widetilde{\phi}_3 \le \frac{456e(1+r_{\max})\log d}{\kappa}.$$

Mixing-window condition. Let

$$t_t := \max\{1, t_{\tau_t}\}, \qquad \tau_{i,j} := \sum_{\ell=i}^{j} \tau_\ell.$$

The mixing-window condition required by Theorem D.3 is

$$\tau_{t-t_t,t-1} \le \chi_{2\mathrm{T}} := \min\left\{\frac{\kappa^2}{1824e(1+r_{\max})\widehat{A}^2\log d}, \frac{1}{4A_0}\right\}. \tag{70}$$

Since the target is $\Delta^\star = 0$, define

$$c_1 := \left(2\|\Delta_0\|_\infty + \frac{B_0}{A_0}\right)^2, \qquad c_2 := \widehat{B}^2,$$

and let

$$K := \min\{t : t \ge t_t\}.$$

Applying Theorem D.3 gives, for every $t \ge K$,

$$\mathbb{E}\|Q_t - Q^\star\|_\infty^2 \le 3c_1 \exp\left\{-\frac{\kappa}{2}\sum_{j=K}^{t-1}\tau_j\right\} + \frac{456e(1+r_{\max})c_2\log d}{\kappa}\sum_{i=K}^{t-1}\tau_i\tau_{i-t_i,i-1}\exp\left\{-\frac{\kappa}{2}\sum_{j=i+1}^{t-1}\tau_j\right\}. \tag{71}$$

Since $A_t - \lambda(V_t) = Q_t - Q^\star$, the same bound holds for $\mathbb{E}\|A_t - \lambda(V_t)\|_\infty^2$.

Theorem D.10 (Finite-time tracking of two-timescale V-SHARE)**.** *Suppose the timescale-separation condition equation 69 and the Markovian mixing-window condition equation 70 hold. Then equation 71 holds for every $t \geq K$.*

In particular, the slow value update enters the finite-time recursion only through the ratio $r_t = \alpha_t/\tau_t$. Once this ratio is uniformly small enough to be absorbed into the dominant fast drift, it changes only the constants of the Markovian finite-time bound and does not generate a separate moving-target error term.

Polynomial gains. Consider

$$\tau_t = \frac{\tau_0}{(t+h)^p}, \qquad \alpha_t = \frac{\alpha_0}{(t+h)^q}, \qquad \frac{1}{2} < p < q \leq 1.$$

Then

$$r_t = \frac{\alpha_0}{\tau_0}(t+h)^{-(q-p)}$$

is non-increasing. Hence the separation condition equation 69 is satisfied whenever

$$\frac{\alpha_0}{\tau_0} h^{-(q-p)} \leq c_{\text{sep}}.$$

For sufficiently large h, the mixing-window condition equation 70 also holds. Moreover, for $t \geq K$,

$$\tau_{t-t_t,t-1} \leq 2t_t\tau_t.$$

Substituting this into equation 71 and applying the standard polynomial-gain unrolling yields

$$\mathbb{E}\|Q_t - Q^\star\|_\infty^2 = O\left(\frac{t_t}{(t+h)^p}\right). \tag{72}$$

Since the Markov chain is geometrically mixing, $t_t = O(\log(t+h))$, and therefore

$$\mathbb{E}\|Q_t - Q^\star\|_\infty^2 = O\left(\frac{\log t}{t^p}\right).$$

The same rate holds for $\mathbb{E}\|A_t - \lambda(V_t)\|_\infty^2$.

Remark 9 (Why no α_t^2/τ_t term appears)**.** *Standard two-timescale finite-time analysis involves a moving target term of order α_t^2/τ_t (Chandak et al., 2025). However our bound does not involve such a term. This is because the identity $A_t - \lambda(V_t) = Q_t - Q^\star$ makes the tracking error autonomous. The slow value update therefore does not move the target of the fast recursion. Instead, it appears as the state-dependent Markovian perturbation $r_t G(\Delta_t, Y_{t+1})$. Once r_t is small enough to be absorbed into the fixed fast Moreau drift, it changes only the constants in the one-step recursion.*

E USEFUL TECHNICAL LEMMAS

Lemma E.1 (Generalized Moreau-envelope properties)**.** *Let*

$$f(x) := \frac{1}{2}\|x\|_c^2, \qquad g(x) := \frac{1}{2}\|x\|_s^2,$$

where g is L-smooth with respect to $\|\cdot\|_s$. Suppose the norms $\|\cdot\|_c$ and $\|\cdot\|_s$ satisfy

$$\ell_{cs}\|x\|_s \leq \|x\|_c \leq u_{cs}\|x\|_s \qquad \forall x \in \mathbb{R}^d,$$

for some $\ell_{cs}, u_{cs} > 0$. For $\theta > 0$, define the generalized Moreau envelope

$$\mathcal{M}_\theta(x) := \min_{u \in \mathbb{R}^d} \left\{ \frac{1}{2}\|u\|_c^2 + \frac{1}{2\theta}\|x-u\|_s^2 \right\}.$$

Then the following hold.

1. *$\mathcal{M}_\theta$ is convex and L/θ-smooth with respect to $\|\cdot\|_s$. In particular, for all x, y,*

$$\mathcal{M}_\theta(y) \leq \mathcal{M}_\theta(x) + \langle \nabla\mathcal{M}_\theta(x), y-x\rangle + \frac{L}{2\theta}\|y-x\|_s^2.$$

2. *The envelope is equivalent to f:*

$$\left(1+\theta l_{cs}^2\right)\mathcal{M}_\theta(x) \le f(x) \le \left(1+\theta u_{cs}^2\right)\mathcal{M}_\theta(x).$$

3. *There exists a norm $\|\cdot\|_m$ such that*

$$\mathcal{M}_\theta(x) = \frac{1}{2}\|x\|_m^2.$$

Moreover, defining

$$\ell_{cm} := \sqrt{1+\theta\ell_{cs}^2}, \qquad u_{cm} := \sqrt{1+\theta u_{cs}^2},$$

we have

$$\ell_{cm}\|x\|_m \le \|x\|_c \le u_{cm}\|x\|_m.$$

4. *Let $\|\cdot\|_m^*$ denote the dual norm of $\|\cdot\|_m$. For every $x \neq 0$ and every $v_x \in \partial\|x\|_m$,*

$$\|v_x\|_m^* \le 1.$$

Proof. Items 1-3 are the generalized Moreau-envelope properties of Chen et al. (2020, Lemma 2.1 and Appendix A.1), rewritten using θ as the smoothing parameter.

For the final claim, the map $x \mapsto \|x\|_m$ is convex and 1-Lipschitz with respect to $\|\cdot\|_m$. Hence, by Chen et al. (2020, Lemma A.1), every $v_x \in \partial\|x\|_m$ satisfies

$$\|v_x\|_m^* \le 1.$$

□

We restate (Chen et al., 2024, Lemma 2.1(2)) for convenience.

Corollary E.2 (ℓ_∞ specialization)**.** *Let $\|\cdot\|_c = \|\cdot\|_\infty$, $\|\cdot\|_s = \|\cdot\|_p$ with $p = 2\log d$. Then one may take $L = p-1 \le 2\log d$, $\ell_{cs} = e^{-1/2}$, and $u_{cs} = 1$. For*

$$\theta := \left(\frac{1+\rho}{2\rho}\right)^2 - 1,$$

the constants above satisfy

$$\phi_1 \le 3, \qquad \phi_2 \ge \frac{1-\rho}{2}, \qquad \phi_3 \le \frac{456e\log d}{1-\rho}.$$

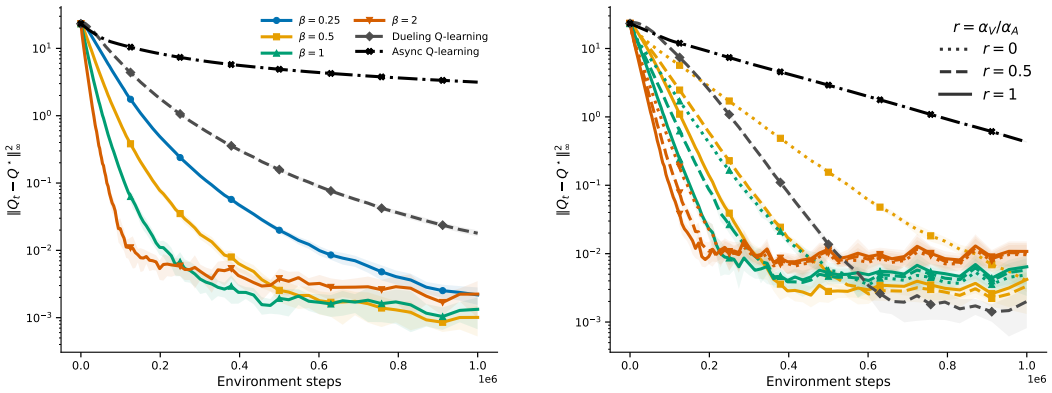
F Experimental Details

F.1 Discrete MDP

We provide further experiments of V-SHARE and it's variants in various settings in Figure 2. Figure 2(b) reveals an interesting separation between the roles of β and the timescale ratio $r := \alpha_V/\alpha_A$. Empirically, β largely determines the terminal noise floor, whereas r primarily controls how rapidly the algorithm approaches that floor. This suggests a potentially useful two-timescale design principle: the advantage stepsize may determine the accuracy ultimately reached, while the relative value stepsize controls the transient speed. A natural direction for future work is to determine whether this separation can be formalized into sharper sample-complexity guarantees, with α_A (or β) controlling the attainable ϵ-accuracy and α_V controlling how quickly that accuracy is reached.

F.2 Deep RL Experiments

Unlike a conventional dueling architecture with a shared feature extractor, our implementation uses separate neural networks for the value and advantage components. This design allows their learning rates to be varied independently, as motivated by the two-timescale analysis in Section 3.5.



(a) Inverse linear stepsize.

(b) Two-timescale V-SHARE

Figure 2: **Tabular experiments on randomly generated MDPs.** (a): performance of single-timescale V-SHARE (Algorithm 1) for different residual weights β under inverse linear stepsizes, compared with asynchronous Q-learning and Dueling Q-learning. (b): comparison of two-timescale V-SHARE on with varying stepsize ratios and β values using constant stepsizes. Curves show the mean squared sup-norm error $\mathbb{E}\|Q_t - Q^\star\|_\infty^2$ over independent trajectories; shaded regions denote pointwise 95% confidence intervals. The y-axis is in the logarithmic scale in all plots.

Classic Control. For Classic Control, each component is represented by a multilayer perceptron (MLP) based on the CleanRL DQN implementation (Huang et al., 2022), with hidden dimensions 120 and 84. Each hidden linear layer is followed by LayerNorm and a ReLU activation. The value network terminates in a scalar output $V(s)$, whereas the advantage network outputs an $|\mathcal{A}|$-dimensional vector $A(s, \cdot)$.

The training horizon is environment dependent: $200,000$ environment steps for Acrobot-v1 and CartPole-v1,$500,000$ steps for MountainCar-v0, and 10^6 steps for LunarLander-v2. Learning starts after $1,000$ initial random-action steps. We use Adam with a base learning rate of 2.5×10^{-4}, discount factor $\gamma = 0.99$, a replay-buffer capacity of 10^4, and minibatches of 128 transitions, with one optimization step every 10 environment steps. The exploration rate is linearly annealed from 1 to 0.05 over the first 50% of training. Unless otherwise specified, we use $\beta = 1$, separate value and advantage networks, and a target network. For the learning-rate-ratio experiments, the value network uses the base learning rate, while the advantage-network learning rate is multiplied by 0.25, 0.5, 1, 2, or 4. For the β-scheduling variant, β is linearly decreased from 100 to 1 over the first 10% of training and remains fixed at 1 thereafter.

We evaluate greedily every 10% of the corresponding training horizon using five fixed evaluation seeds, $0, 1, 2, 3, 4$, with one episode per evaluation seed. Results are aggregated over five independent training seeds, $1, \ldots, 5$.

MinAtar. For MinAtar, each network consists of a convolutional layer with 16 (3×3) filters and stride 1, followed by a ReLU activation, a fully connected layer with 128 hidden units, and another ReLU activation. The advantage network terminates in a linear layer with $(|\mathcal{A}|)$ outputs, whereas the value network terminates in a linear layer with a single scalar output.

Experience is stored in a replay buffer with capacity 10^6 transitions. Learning begins after the first 80,000 environment steps, after which the agent samples a minibatch of 32 transitions from the replay buffer every four environment steps. Both the value and advantage networks are optimized independently using Adam with a base learning rate of 10^{-4}. During training, actions are selected using an ϵ-greedy policy. Exploration begins with $\epsilon = 1$ and is linearly annealed to 0.01 over the first 10% of training, corresponding to the first one million environment steps for a 10-million-step run. Thereafter, ϵ remains fixed at 0.01.

We periodically evaluate the learned policy every 200,000 environment steps. We evaluate greedily using five fixed evaluation seeds, $0, 1, 2, 3, 4$, running one episode per seed at each evaluation point. We use no exploratory actions during evaluation.

For both Classic Control and MinAtar, the two components are combined as

$$Q(s, a; \theta_V, \theta_A) = V(s; \theta_V) + A(s, a; \theta_A).$$

For a sampled transition $(S, A, R, S', \pi_b(A \mid S))$, the bootstrap target is:

$$Y = R + \gamma \left[V(S') + \max_{a'} A(S', a') \right].$$

The corresponding temporal-difference (TD) residual is

$$\delta = Y - [V(S) + A(S, A)],$$

and the TD objective is given by

$$L_{\mathrm{TD}} = \mathbb{E}\left[(Y - V(S) - A(S, A))^2 \right].$$

Gradients from L_{TD} are not propagated to θ_A, since the advantage network is updated separately using the preconditioned update rule.

In the deep RL setting, individual updates cannot be directly controlled, and the inverse behavior probability $1/\pi_b(A \mid S)$ can become large for infrequently selected actions. Such large updates may adversely affect training stability. We set $\beta = 1$ by default. For Classic Control, we use the residual-preconditioning weight $w = \frac{1}{\pi_b(A|S)}$ as large inverse behavior probabilities were not a practical issue in these environments. For MinAtar, we clip the inverse behavior probability by defining

$$\rho(S, A) = \min \left\{ \frac{1}{\pi_b(A \mid S)}, \rho_{\max} \right\},$$

where $\rho_{\max} = 5$. The residual-preconditioning weight is then defined as

$$w = \beta \rho(S, A) = \rho(S, A).$$

Using this weight, we construct the auxiliary advantage target

$$\hat{A}(S, A) = A(S, A) + w\delta.$$

The corresponding advantage loss is

$$L_{\mathrm{adv}} = \frac{1}{2} \mathbb{E}\left[\left(A(S, A) - \hat{A}(S, A) \right)^2 \right].$$

In addition to the TD and advantage objectives, we apply an explicit regularization penalty to both components of the action-value decomposition. This regularization is introduced to address the identifiability issue that arises in advantage–value decomposition methods (Daley et al., 2025; Wang et al., 2016). Specifically, we use

$$L_{\mathrm{reg}} = \frac{\lambda}{2} \mathbb{E}\left[V(S)^2 + \sum_{a \in \mathcal{A}} A(S, a)^2 \right],$$

where $\lambda = 5 \times 10^{-3}$, we did not optimize for the best λ.

The complete training objective is therefore

$$L = L_{\mathrm{TD}} + L_{\mathrm{adv}} + L_{\mathrm{reg}}.$$

Importantly, although these terms are written as a single objective, gradients from L_{TD} are stopped with respect to the advantage-network parameters θ_A; consequently, θ_A is updated through L_{adv} and L_{reg}, while the TD loss directly updates the value-network parameters.

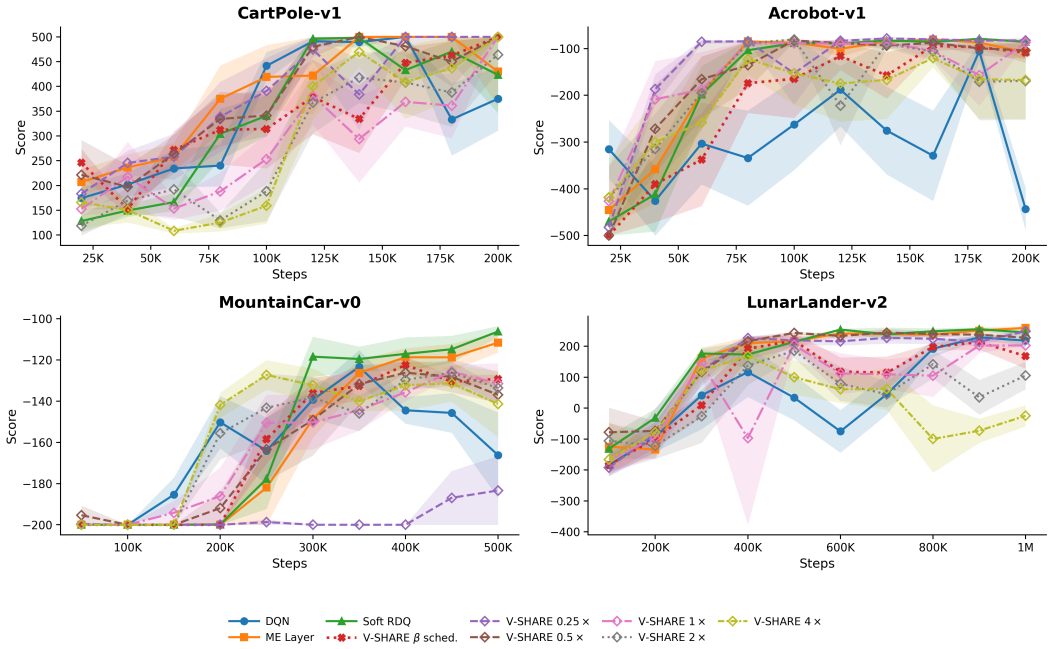


Figure 3: Learning curves for DQN, ME Layer, Soft RDQ, and V-SHARE across four classic-control environments. We evaluate V-SHARE with different advantage learning-rate multipliers and with decreasing-β scheduling to study the effects of timescale separation and residual-preconditioning schedules in the deep RL setting. Curves show the mean evaluation return over five independent training seeds, with shaded regions indicating ± 1 SEM.

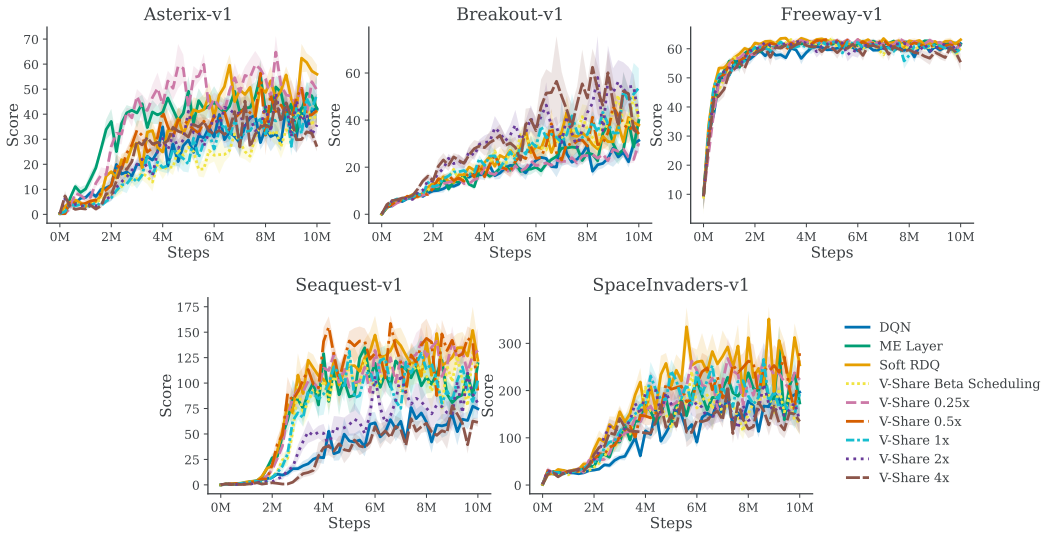


Figure 4: Learning curves for DQN, ME Layer, Soft RDQ, and V-SHARE across five MinAtar environments. We evaluate V-SHARE with different advantage learning-rate multipliers and with decreasing-β scheduling to study the effects of timescale separation and residual-preconditioning schedules in the deep RL setting. Curves show the mean evaluation return over six independent training seeds, with shaded regions indicating ± 1 SEM.