RANDOMIZED LEARNING-RATE Q-LEARNING FOR CONTINUING DISCOUNTED MDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

Randomized learning-rate Q-learning achieves tractable, gap-independent regret in episodic finite-horizon MDPs and often outperforms optimistic-bonus Q-learning, but its guarantees rely on a terminal boundary that ends the Bellman recursion. Removing this boundary leads to continuing discounted control, for which no such guarantee is known. We propose a discounted randomized Q-learning algorithm (D-RandQ), whose analysis lifts stage-wise reuse counting to a global interaction-time occurrence representation: each target is aligned with its successor occurrence, stage-ending self-loops and boundary visits are charged separately, and the recursion closes through a reuse operator with bounded column sum. Combined with adaptive-update concentration and Dirichlet anti-concentration, this yields a finite-sample, high-probability, gap-independent action-gap bound, hence, to our knowledge, the first finite-sample, gap-independent sublinear regret guarantee for randomized learning-rate Q-learning in continuing discounted MDPs. Empirically, under a documented disjoint calibration protocol, D-RandQ attains lower cumulative action gaps and higher realized rewards than Q-SlowSwitch-Adv and UCBVI-γ on Maze-25, and a large separation on the reported Chain-50 runs at the same transferred operating point.

1 INTRODUCTION

Model-free reinforcement learning (RL) learns value estimates without planning with an explicit transition model (Sutton, 1988; Watkins & Dayan, 1992). The simplest regime is episodic interaction in a finite-horizon Markov decision process (MDP), where a terminal boundary closes the Bellman recursion; there, randomized learning-rate Q-learning, which randomizes the scalar learning rate of Q-updates rather than sampling entire value functions, gives a computationally tractable, gap-independent regret guarantee (Tiapkin et al., 2023; Wang et al., 2026). Removing the reset makes the problem strictly harder, since the interaction never terminates and no boundary remains. Continuing control has been studied along two threads: average-reward control (Wei et al., 2020) and discounted control, whose regret has been formalized through several criteria, namely reward-based γ-regret (Liu & Su, 2020), the value regret of UCBVI-γ (He et al., 2020), and the frozen-policy value regret of Q-SlowSwitch-Adv (Ji & Li, 2023).

However, no randomized learning-rate Q-learning algorithm and regret guarantee are currently known for continuing discounted MDPs: the analyses of RandQL (Tiapkin et al., 2023) and RandomizedQ (Wang et al., 2026) are episodic and rely on a terminal boundary that this setting does not provide. We therefore ask: ***Can a Q-learning algorithm employ randomized learning-rate exploration in a continuing discounted MDP while attaining gap-independent sublinear regret?***

1.1 MAIN CONTRIBUTIONS

In the continuing discounted setting, we study finite tabular MDPs with S states and A actions, rewards in $[0, 1]$, and a known discount factor $\gamma \in (0, 1)$, over T interactions, without resets, ergodicity, or a minimum-gap assumption. Our contributions are three-fold.

- **Algorithm design.** We propose a discounted randomized Q-learning algorithm for continuing discounted MDPs. It inherits RandomizedQ's beta learning rates, fast/slow ensembles,

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.

and per-step greedy updates (Wang et al., 2026), and adds a stage schedule with an initial stage length L and growing stage lengths $M_q = \lfloor L(1 + 1/L)^q \rfloor$, together with mixing coefficients $\eta_q = (M_q + 1)^{-1/2}$. The slow ensemble commits at stage boundaries and supplies the optimism used in the proof, while the fast heads absorb each new transition so the greedy policy can change every step; both targets are frozen before the beta draws. Each transition costs $O(A + J)$ time and the algorithm uses $O(JSA)$ memory for J heads.

- **Theoretical analysis.** We give the first continuing discounted regret analysis for this line, centered on lifting prior stage-wise reuse counting to a global interaction-time occurrence representation: each read is aligned with the correct value version, and stage-ending self-loops, first post-commit visits, and boundary occurrences are charged separately. A reuse operator of column sum at most $1 + 2/L$ closes the discounted recursion, and matrix powers track further propagation through the changing fast/slow mixture (Lemmas 16 and 17). This yields a finite-sample, gap-independent action-gap bound that holds with probability at least $1 - \delta$ (Theorem 1), which is $O_\gamma(T/\sqrt{\log T})$ for fixed S, A, γ, δ, plus an exact Bellman-telescoping conversion to Liu–Su reward-based γ-regret (Corollary 1). The current analysis does not capture a theoretical agility benefit from the fast path: its bootstrap contribution is bounded by its global range and charged non-recursively.
- **Experimental validation.** On Maze-25, under practical parameters selected on a disjoint calibration task, D-RandQ attains lower cumulative action gap and higher cumulative reward than model-free Q-SlowSwitch-Adv and model-based UCBVI-γ on every paired base seed. On Chain-50, at the same transferred operating point, it shows a large empirical separation as well, although separate calibration artifacts indicate that the optimistic baselines are sensitive to bonus scaling, so we do not interpret the Chain comparison as tuning-independent dominance. Ablations show that the full fast/per-step architecture substantially outperforms a committed slow-estimate variant, which jointly removes the fast heads and mixture, changes the policy-refresh schedule, and changes the bootstrap source from $\widehat{V}$ to $\widetilde{V}^{\flat}$; the comparison is therefore architectural and does not isolate one mechanism (Section 5).

1.2 Related work

Table 1 summarizes the closest model-free sampling-based methods; we position our contribution along three lines of work.

Key property	**Conditional-PS** (Dann et al., 2021)	**Staged-RandQL** (Tiapkin et al., 2023)	**RandQL** (Tiapkin et al., 2023)	**RandomizedQ** (Wang et al., 2026)	**This Work**
Computationally tractable	×	✓	✓	✓	✓
Per-step policy updates	✓	×	✓	✓	✓
Continuing discounted action-gap guarantee	×	×	×	×	✓

Table 1: Structural comparison of representative model-free sampling-based methods in tabular RL. The first four methods are analyzed in episodic settings, whereas our displayed continuing guarantee concerns the action-gap criterion of Section 2. The table does not compare numerical regret rates proved under different criteria.

Optimistic Q-learning and continuing control. Optimistic Q-learning anchors the finite-sample theory of tabular model-free control, spanning PAC guarantees (Strehl et al., 2006), episodic regret (Jin et al., 2018; Zhang et al., 2020; Li et al., 2021), and gap-dependent regret under a positive minimum gap (Simchowitz & Jamieson, 2019; Yang et al., 2021). Dropping episodic resets has been pursued for average-reward learning under weak communication or ergodicity (Wei et al., 2020). For discounted MDPs, reward-based γ-regret bounds (Liu & Su, 2020), model-based UCBVI-γ with Bernstein bonuses (He et al., 2020), and model-free Q-SlowSwitch-Adv with variance reduction and adaptive switching (Ji & Li, 2023) each provide continuing guarantees. Classical online Q-learning without optimism has also received high-probability discounted continuing analysis, including a gap-robust $\widetilde{O}(T^{9/10})$ result for smoothed ϵ_t-greedy exploration (Singh et al., 2026); that method does not randomize learning rates or use our continuing occurrence-time analysis. UCBVI-γ uses a non-stationary value-regret criterion, whereas Q-SlowSwitch-Adv primarily uses frozen-policy value regret; we do not compare our original action-gap rate numerically with either value-regret rate.

Throughout this comparison, γ-regret denotes the reward-based criterion of Liu & Su (2020), explicitly defined in Section 4. Under that criterion, the minimax leading scale is $\widetilde{O}(H_\gamma^{1/2}\sqrt{SAT})$ with $H_\gamma = (1-\gamma)^{-1}$ (He et al., 2020; Zamir et al., 2026); the native $\widetilde{O}(H_\gamma^{3/2}\sqrt{SAT})$ result of He et al. (2020) uses a distinct non-stationary value-regret normalization, and their Appendix A.2 comparison introduces an approximate factor $1-\gamma = H_\gamma^{-1}$ up to endpoint and martingale terms, so the converted leading scale should not be read as their complete finite-T theorem, whose regime and lower-order terms still matter. Zamir et al. (2026) treat Liu–Su-style γ-regret directly: their optimal variance-dependent guarantee retains that minimax worst-case scale while adapting to variance/span and improving on benign instances. Our fixed-discount bound is sublinear in T but does not claim minimax optimality in the discount factor, and the criteria remain incomparable; Appendix I gives a compact criterion-aware comparison.

Randomized exploration and learning-rate randomization. Randomized exploration replaces optimism with sampled estimates. Posterior sampling admits tabular regret guarantees (Osband et al., 2013; Agrawal & Jia, 2017), including non-episodic formulations (Osband & Van Roy, 2016); scalable practice randomizes ensembles or network parameters (Osband et al., 2016; Fortunato et al., 2018); and randomized least-squares value iteration randomizes value estimates with frequentist guarantees (Zanette et al., 2020). These schemes differ from scalar learning-rate randomization in what they sample and in the computation required. The closest line randomizes the learning rate itself. Dann et al. (2021) give model-free episodic posterior-sampling guarantees, although sampling value functions directly can be computationally demanding. Tiapkin et al. (2023) introduce the tractable beta-learning-rate RandQL construction with step-wise value updates, while their tabular regret guarantee is established for Staged-RandQL, which commits policy values only at stage boundaries. RandomizedQ combines fast/slow ensembles with beta learning rates to obtain a certified episodic guarantee while retaining per-step greedy adaptation (Wang et al., 2026). Wang et al. (2026) also identify an aggregated-weight partial-sum martingale issue in the earlier Staged-RandQL analysis: the asserted filtration cannot simultaneously make those partial sums adapted and future aggregate weights independent. We do not reuse that step: our targets are frozen before the current beta draws, so randomized-rate errors expand with deterministic mean-weight backward coefficients and are controlled by conditional-MGF and stopped calendar-time martingales. D-RandQ retains RandomizedQ's fast/slow architecture; our contribution is to lift that reuse accounting to global interaction-time occurrences, where value versions must be aligned and stage-ending self-loops require separate charges. This representation closes the discounted recursion without a terminal Bellman layer; it does not claim the column-sum idea itself as new.

2 Problem setup

Discounted MDP. Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$ be a finite discounted MDP with nonempty state and action sets, $|\mathcal{S}| = S \geq 1$, $|\mathcal{A}| = A \geq 1$, $\gamma \in (0,1)$, and $H_\gamma = (1-\gamma)^{-1}$. Conditional on the history before the transition at time t, $R_t \in [0,1]$ has mean $r(s_t, a_t)$ and $s_{t+1} \sim P(\cdot \mid s_t, a_t)$. Rewards and next states may be coupled within one transition. For a pair $x = (s,a)$, write $P_x f := \sum_{s' \in \mathcal{S}} P(s' \mid s,a) f(s')$ for the transition-kernel expectation of a function f. The initial state s_1 is either fixed or sampled from a prescribed distribution before the first transition; all bounds below hold conditionally on any realized s_1. We impose no ergodicity, reset, or minimum-visitation assumption. The conditional kernel specified below governs each transition; the policy and learning-rate parameters may depend on the preceding history.

Interaction protocol. Let $\mathcal{F}_t$ be the complete history immediately before action selection. The state, stored tables, and action selected using a fixed deterministic tie-breaking rule are $\mathcal{F}_t$-measurable. Conditional on this history, (R_t, s_{t+1}) has the fixed joint kernel K_{s_t,a_t}. After observing the transition, the algorithm freezes both targets and then draws its fresh beta learning rates, independently across heads and ensembles. Appendix A.2 gives the filtration and randomness conventions used in the proofs.

Optimal values. All logarithms below are natural logarithms. The optimal Bellman equations are

$$Q^*(s,a) = r(s,a) + \gamma P_{s,a} V^*, \qquad V^*(s) = \max_a Q^*(s,a), \qquad 0 \leq Q^*, V^* \leq H_\gamma.$$

Action-gap regret. Define the one-step optimal-action gap and its cumulative regret by

$$g_t = V^*(s_t) - Q^*(s_t, a_t), \qquad \mathrm{Regret}_\Delta(T) = \sum_{t=1}^{T} g_t. \tag{1}$$

Thus g_t measures the loss of the selected action relative to an optimal action at the current state, with optimal behavior used after that action; through Q^* it includes the discounted downstream effect of that decision. Our goal is to control $\mathrm{Regret}_\Delta(T)$ without assuming a positive minimum action gap. This criterion evaluates decisions along the realized state sequence; it does not compare the agent's cumulative realized reward with that of an optimal policy following a different state sequence. For instance, if an initial zero-reward action leads irreversibly either to a unit-reward or to a zero-reward absorbing state and the learner selects the latter, then $\mathrm{Regret}_\Delta(T) = \gamma/(1-\gamma)$, whereas the counterfactual undiscounted finite-horizon cumulative reward difference is $R_{\mathrm{cf}}(T) = T - 1$. This example separates these two quantities, but not from every discounted-regret or value-regret definition; in particular, Lemma 1 converts to Liu–Su γ-regret. The distinction matters when no communication or recurrence assumption is imposed.

Horizon and snapshot convention. The analysis concerns the first T transitions. A subscript t on any stored table denotes its complete snapshot immediately before a_t is selected, and an update after transition t first appears at time $t + 1$. Both TD targets use the pre-update tables, including for a self-loop, and are frozen before the learning rates are sampled. Unmodified entries retain their previous values, so the value read by a target is distinguished from the value available at the next action selection.

3 The Discounted Randomized Q-Learning algorithm

Horizon awareness. The finite-sample algorithm and theorem are horizon-aware: the interaction budget T is assumed known in advance and sets the theoretical parameters (L and the stage lengths in (2), and ℓ, J, $\kappa^\flat$, and $n_0^\flat$ in (5)), so the stated result is not an anytime guarantee. For fixed γ, a sufficiently large constant L would already enforce contraction, but the growth $L = L(T)$ used here is also what makes $\bar{\eta} = (L+1)^{-1/2}$ vanish and hence turns the uniform $O(\bar{\eta}T)$ mixture charge into a sublinear term. An anytime version, via an epoch-based or time-varying parameter schedule, is outside the scope of the present theorem.

Stage schedule and theoretical parameters. All logarithms are natural. The integer L controls the initial stage length and subsequent stage growth. Set

$$U = 2H_\gamma, \qquad L = \max\left\{1, \left\lceil \frac{\log(TH_\gamma^2)}{\log(1/\gamma)} \right\rceil \right\}, \qquad M_q = \max\{1, \lfloor L(1+1/L)^q \rfloor\}. \tag{2}$$

The mixing rate in slow stage q is $\eta_q = (M_q + 1)^{-1/2}$, and

$$\bar{\eta} = (L+1)^{-1/2}, \qquad \eta_q \le \bar{\eta}. \tag{3}$$

When a stored policy entry is read, $\eta_t(s,a)$ denotes the rate used in its most recent policy-Q update. Before the first update of a pair we set $\eta_t(s,a) = \eta_0 := (M_0+1)^{-1/2}$, where $M_0 = L$; in every case it is at most $\bar{\eta}$. We use the selected-pair abbreviation

$$\eta_t := \eta_t(s_t, a_t) \tag{4}$$

only when the arguments are exactly (s_t, a_t); rates for comparison actions retain both arguments. For the theorem, let

$$\ell = \log \frac{128SAT}{\delta}, \qquad J = \lceil c_J \ell \rceil, \qquad \kappa = 1, \quad n_0 = 1, \qquad \kappa^\flat = c_\flat \ell, \quad n_0^\flat = \lceil c_0 \kappa^\flat \ell \rceil, \tag{5}$$

where $c_J, c_\flat, c_0$ are fixed universal constants chosen as in Appendix A.2.1. In particular, these constants do not depend on the MDP, horizon, discount, or confidence level. Throughout the proofs, C denotes a universal constant that may change between displays.

Algorithm 1 Discounted Randomized Q-Learning

Require: initial state s_1, horizon T, discount γ, confidence δ, and the parameters $J, \kappa, \kappa^\flat, n_0, n_0^\flat$ from (5)

$L \leftarrow \max\{1, \lceil \log(TH_\gamma^2)/\log(1/\gamma) \rceil\}$, $U \leftarrow 2H_\gamma$
Initialize $N(s,a) = N^\flat(s,a) = q(s,a) = 0$ for every (s,a)
Initialize $\widehat{Q}^j = \widehat{Q}^{\flat,j} = \widetilde{Q}^\flat = Q = U$ for all $(s,a), j$; set $\widehat{V}(s) = \widetilde{V}^\flat(s) = U$ for every s
for $t = 1, \ldots, T$ **do**
  $a_t \in \arg\max_a Q(s_t, a)$; observe R_t and s_{t+1}
  $m \leftarrow N(s_t, a_t)$, $m^\flat \leftarrow N^\flat(s_t, a_t)$, $q_0 \leftarrow q(s_t, a_t)$
  $M \leftarrow \max\{1, \lfloor L(1 + 1/L)^{q_0} \rfloor\}$, $\eta \leftarrow (M+1)^{-1/2}$
  Freeze $Y_t \leftarrow R_t + \gamma \widehat{V}(s_{t+1})$ and $Y_t^\flat \leftarrow R_t + \gamma \widetilde{V}^\flat(s_{t+1})$
  for $j = 1, \ldots, J$ **do**
    Sample $w_m^j \sim \mathrm{Beta}((L+1)/\kappa, (m+n_0)/\kappa)$
    $\widehat{Q}^j(s_t, a_t) \leftarrow (1 - w_m^j)\widehat{Q}^j(s_t, a_t) + w_m^j Y_t$
    Sample $w_{m^\flat}^{\flat,j} \sim \mathrm{Beta}(1/\kappa^\flat, (m^\flat + n_0^\flat)/\kappa^\flat)$
    $\widehat{Q}^{\flat,j}(s_t, a_t) \leftarrow (1 - w_{m^\flat}^{\flat,j})\widehat{Q}^{\flat,j}(s_t, a_t) + w_{m^\flat}^{\flat,j} Y_t^\flat$
  $Q(s_t, a_t) \leftarrow \eta \max_j \widehat{Q}^j(s_t, a_t) + (1 - \eta)\widetilde{Q}^\flat(s_t, a_t)$
  $\pi(s_t) \in \arg\max_a Q(s_t, a)$; $\widehat{V}(s_t) \leftarrow \max_j \widehat{Q}^j(s_t, \pi(s_t))$
  $N(s_t, a_t) \leftarrow m + 1$; $N^\flat(s_t, a_t) \leftarrow m^\flat + 1$
  if $N^\flat(s_t, a_t) = M$ **then**
    $\widetilde{Q}^\flat(s_t, a_t) \leftarrow \max_j \widehat{Q}^{\flat,j}(s_t, a_t)$
    $Q(s_t, a_t) \leftarrow \eta \max_j \widehat{Q}^j(s_t, a_t) + (1 - \eta)\widetilde{Q}^\flat(s_t, a_t)$
    $\pi(s_t) \in \arg\max_a Q(s_t, a)$; $\widehat{V}(s_t) \leftarrow \max_j \widehat{Q}^j(s_t, \pi(s_t))$
    $\pi^\flat(s_t) \in \arg\max_a \widetilde{Q}^\flat(s_t, a)$; $\widetilde{V}^\flat(s_t) \leftarrow \widetilde{Q}^\flat(s_t, \pi^\flat(s_t))$
    $q(s_t, a_t) \leftarrow q_0 + 1$; $N^\flat(s_t, a_t) \leftarrow 0$; $\widehat{Q}^{\flat,j}(s_t, a_t) \leftarrow U$ for every j

Stored estimates and updates. The algorithm maintains J fast heads $\widehat{Q}^j$, J temporary slow heads $\widehat{Q}^{\flat,j}$, a committed slow table $\widetilde{Q}^\flat$, and a mixture Q used for greedy action selection. Fast heads and their cumulative counter $N(s,a)$ persist throughout the run. Each slow stage starts with temporary heads equal to U and a local counter $N^\flat(s,a) = 0$. At a stage boundary, the maximum temporary slow head is committed, and only the temporary slow heads and their local counter are reset.

Both ensembles use targets frozen before any update, and their maxima are evaluated after all heads have been updated. A commit immediately synchronizes the affected policy entry and state values using the completed stage's mixing rate, while the next rate is installed on the next visit to that pair; this first post-commit visit is charged separately in the analysis. All other table entries remain unchanged, hats denote temporary or fast estimates, and tildes denote committed slow estimates.

Roles of the two ensembles. As inherited from RandomizedQ, the fast heads incorporate each new transition immediately, allowing the greedy policy to change between slow commits, while the slow ensemble provides the optimism used in the proof. Its temporary heads remain hidden from action selection and both targets until commit, which is essential to the conditional-independence argument. The mixing coefficient $\eta_q = (M_q + 1)^{-1/2}$ decreases as stages grow, limiting the fast component's contribution to estimation error. Our present upper-bound analysis treats the fast path as a bounded perturbation, so any theoretical benefit from its per-step adaptation is not reflected in the resulting regret bound.

Computational cost. Each transition updates $2J$ scalar heads. Because the mixed policy table is stored explicitly, its greedy action costs $O(A)$ and each ensemble maximum for the visited entry costs $O(J)$, so the per-transition time is $O(A + J)$. The memory footprint is $O(JSA)$, and stage counters and the committed tables require only $O(SA)$ additional memory. No model or transition kernel is stored by the algorithm.

Synchronization convention. At every action-selection point, an initialized policy entry has the mixture form

$$Q_t(s,a) = \eta_t(s,a) \max_j \widehat{Q}_t^j(s,a) + (1 - \eta_t(s,a))\widetilde{Q}_t^\flat(s,a), \qquad \eta_t(s,a) \le \bar{\eta},$$

while an unvisited entry equals U. This follows by induction over Algorithm 1: after a stage-ending commit, the algorithm immediately recomputes the affected policy entry and its greedy fast value, so each entry uses the current committed slow value. It may retain the completed stage's mixing rate until the next owner visit, a boundary case charged separately below.

4 MAIN RESULTS

We state the high-probability action-gap bound for D-RandQ and then convert it to the reward-based Liu–Su criterion. For comparison under a reward-based criterion, define Liu–Su γ-regret in our indexing by

$$\mathrm{Regret}_\gamma^{\mathrm{LS}}(T) = \sum_{t=1}^{T} \{(1-\gamma)V^*(s_t) - R_t\}. \tag{6}$$

Here $(1-\gamma)V^*(s_t)$ is the constant per-step reward whose discounted sum equals the optimal value $V^*(s_t)$, so $\mathrm{Regret}_\gamma^{\mathrm{LS}}(T)$ sums the per-step reward shortfall against the optimal value along the realized trajectory, without any reset. This reward-based criterion is the one under which prior discounted-MDP regret bounds are stated (Liu & Su, 2020; He et al., 2020; Zamir et al., 2026), and it is the natural reward-based counterpart of our action-gap quantity. Lemma 1 in the appendix gives an exact pathwise Bellman-telescoping identity between (6) and $\mathrm{Regret}_\Delta(T)$, including stochastic reward and transition noise.

Theorem 1 (D-RandQ high-probability action-gap regret bound)**.** *For $\delta \in (0,1)$ and an integer $T \geq 2$, run Algorithm 1 with the tuple (T, γ, δ) and the parameters in* (2)–(5)*, the confidence parameter being the same δ as in the probability guarantee. Define $\rho_{\gamma,L} = \gamma(1+2/L)$ and $\Gamma_T = \sum_{k=0}^{T-1} \rho_{\gamma,L}^k$. For the choice of L in* (2)*, one has $0 < \rho_{\gamma,L} < 1$ for every $T \geq 2$ and $\gamma \in (0,1)$, as verified in Appendix B, and therefore $\Gamma_T = (1-\rho_{\gamma,L}^T)/(1-\rho_{\gamma,L}) \leq (1-\rho_{\gamma,L})^{-1}$; we retain the exact finite sum, which is no larger than this infinite-series upper bound. There are universal constants $c_J, c_\flat, c_0, C$ such that, with probability at least $1-\delta$ over both the MDP observations and all randomized learning rates,*

$$\mathrm{Regret}_\Delta(T) \leq \min\left\{ H_\gamma T,\ C\Gamma_T H_\gamma \left[\frac{T}{\sqrt{L+1}} + L\{\ell^2\sqrt{SAT} + SA\ell^3\} \right] \right\}. \tag{7}$$

Proof. Please refer to Appendix B for the proofs of Theorem 1 and of the fixed-discount asymptotics remark. □

Remark. Theorem 1 controls the cumulative optimal-action gaps along the realized trajectory with high probability, without requiring a positive minimum gap. The factor Γ_T quantifies how estimation errors propagate through repeated discounted bootstrapping; inside the brackets, $T/\sqrt{L+1}$ comes from the $O(\bar{\eta}T)$ fast–slow mixture charge, and the remaining terms collect concentration, prior, and stage-boundary contributions. Two effects are worth separating. First, the fast bootstrap is bounded through its global range, $\widehat{V}_i(s) - V^*(s) \leq U = 2H_\gamma$, giving a per-row term proportional to $2H_\gamma\eta_t$ that is absorbed into the nonrecursive charge r_t and controlled, over nonexceptional rows, by the stagewise bound $\sum_{t\ \mathrm{nonexceptional}} \eta_t \leq CL\sqrt{SAT}$ (exceptional rows are charged separately); this range-bounded term is therefore not by itself the direct source of the leading $T/\sqrt{L}$ rate. Second, that rate is generated directly by the uniform approximate-optimism and mixture shift at $\bar{\eta} = (L+1)^{-1/2}$, which yields the per-step drift $3\bar{\eta}H_\gamma$ in the slow-error conversion and master recursion, hence a contribution proportional to $\bar{\eta}T = T/\sqrt{L+1}$; this is $O_\gamma(T/\sqrt{\log T})$ for fixed γ because $L = \Theta_\gamma(\log T)$. Sharpening the coupling between the fast path and the recursive error could reduce this uniform mixture drift, and the deterministic bound $H_\gamma T$ applies at short horizons, making the guarantee meaningful for every finite T.

Remark (Fixed-discount asymptotics). Fix $\delta, \gamma \in (0,1)$. For $T \geq T_0(\gamma)$ one has $\rho_{\gamma,L} \leq (1+\gamma)/2$ and hence $\Gamma_T \leq 2H_\gamma$, so (7) yields a deterministic deviation bound $B_T(\delta)$, explicit in Appendix B, with $\mathbb{P}\{\mathrm{Regret}_\Delta(T) \leq B_T(\delta)\} \geq 1-\delta$, leading term $O_\gamma(T/\sqrt{\log T})$, and $B_T(\delta) = o(T)$ for fixed S, A, γ, δ. Consequently, for the algorithm sequence indexed by its own horizon and the confidence schedule $\delta_T = 1/T$, the average gap $\mathrm{Regret}_\Delta(T)/T$ vanishes in probability, and so does $N_\epsilon(T)/T$ for every fixed $\epsilon > 0$, where $N_\epsilon(T) = \sum_{t=1}^{T} \mathbf{1}\{g_t \geq \epsilon\}$. Both statements are per fixed ϵ and

pointwise in T: no uniform-in-ϵ, almost-sure, or simultaneous-in-T claim is made, and neither implies convergence of the policy on unvisited states.

Corollary 1 (D-RandQ under Liu–Su γ-regret)**.** *Fix $\delta \in (0,1)$ and run Algorithm 1 with confidence parameter $\delta/2$. Put*

$$\ell_{\mathrm{LS}} = \log \frac{256SAT}{\delta}.$$

Then, with probability at least $1-\delta$,

$$\mathrm{Regret}_{\gamma}^{\mathrm{LS}}(T) \le \min\left\{T,\ C\Gamma_T H_\gamma \left[\frac{T}{\sqrt{L+1}} + L\{\ell_{\mathrm{LS}}^2\sqrt{SAT} + SA\ell_{\mathrm{LS}}^3\}\right] + \gamma H_\gamma + H_\gamma\sqrt{\frac{T}{2}\log\frac{2}{\delta}}\right\}. \tag{8}$$

For fixed S, A, γ, δ, the right-hand side is

$$O_\gamma\left(\frac{T}{\sqrt{\log T}}\right) + O_\gamma(\sqrt{T}\,\log^3 T) + O_\gamma(\log^4 T) = o(T).$$

Proof. Please refer to Lemma 1 in Appendix C for the proof. □

Remark. This corollary transfers the action-gap guarantee of Theorem 1 to the reward-based γ-regret criterion of Liu & Su (2020), the natural measure for continuing discounted control. The two additive lower-order terms arise from the Bellman-telescoping boundary (γH_γ) and from the transition-noise martingale tail ($H_\gamma\sqrt{\frac{T}{2}\log\frac{2}{\delta}}$); neither affects the leading rate. The deterministic cutoff T is the pathwise bound for the Liu–Su criterion, not the action-gap cutoff $H_\gamma T$ of Theorem 1.

4.1 Overview of the approach

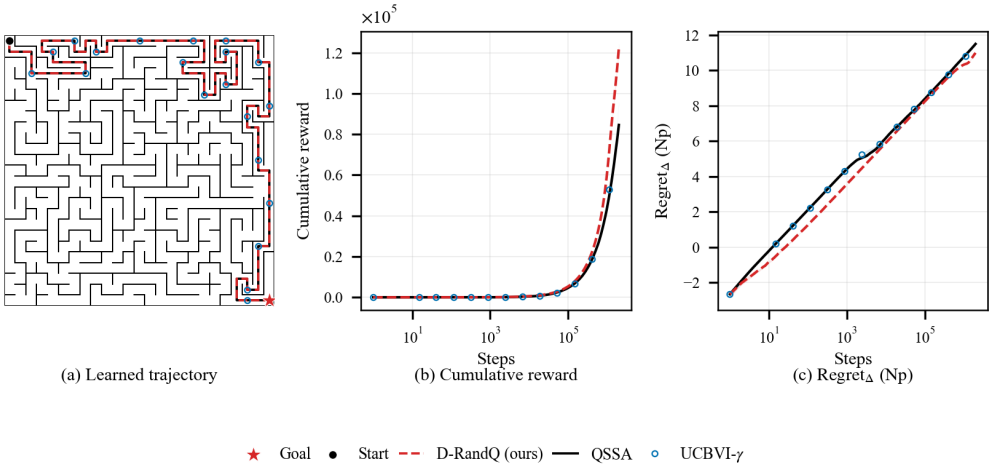
We now explain at a high level how D-RandQ obtains its continuing discounted guarantee and where the novelty lies; a complete roadmap is given at the start of Appendix B. The design separates two roles: the slow ensemble commits its best temporary estimate only at stage boundaries and is never read before a commit, so with high probability its committed values dominate the optimal values, while the fast heads absorb every transition immediately, so the greedy policy can adapt between commits. The proof therefore reads optimism from the slow ensemble and charges the fast ensemble only a bounded extra cost, decoupling adaptivity from statistical stability.

The hard part is that continuing interaction reuses bootstrapped targets indefinitely through cycles and self-loops, so the layer-by-layer counting that closes episodic analyses has no terminal boundary to stop at. Our innovation is to lift this counting to a global interaction-time occurrence representation, in which every read is matched to the exact value version its target used, the few places where the version genuinely changes (stage-ending self-loops, first post-commit visits, and boundary occurrences) are charged separately, and what remains is bounded by a reuse matrix $\boldsymbol{A}$ of column sum at most $1+2/L$. The shifted error vector then obeys $\boldsymbol{d} \le \boldsymbol{r} + \gamma\boldsymbol{A}\boldsymbol{d}$, with $\boldsymbol{r}$ collecting the non-recursive charges; since $\boldsymbol{A}$ is strictly lower triangular and L is chosen so that $\gamma(1+2/L) < 1$, substituting the recursion at most T times yields the geometric series Γ_T and hence the sublinear bound of Theorem 1. The column bound limits how often any single error can be re-read, while strict triangularity forces time forward, so the recursion cannot blow up without a terminal Bellman layer.

5 Experiments

We evaluate practical D-RandQ and Slow-only on Maze-25 and deceptive Chain-50, reporting cumulative action-gap regret and realized reward. Following Wang et al. (2026), we also show representative maze trajectories.

Baselines and protocol. Baselines are model-free QSSA (Q-SlowSwitch-Adv) (Ji & Li, 2023) and a model-based UCBVI-γ (He et al., 2020) baseline used as a practical implementation: it updates its empirical model after every interaction but plans lazily, at count-doubling events, so it is not an



(a) Learned trajectory (b) Cumulative reward (c) Regret$_\Delta$ (Np)

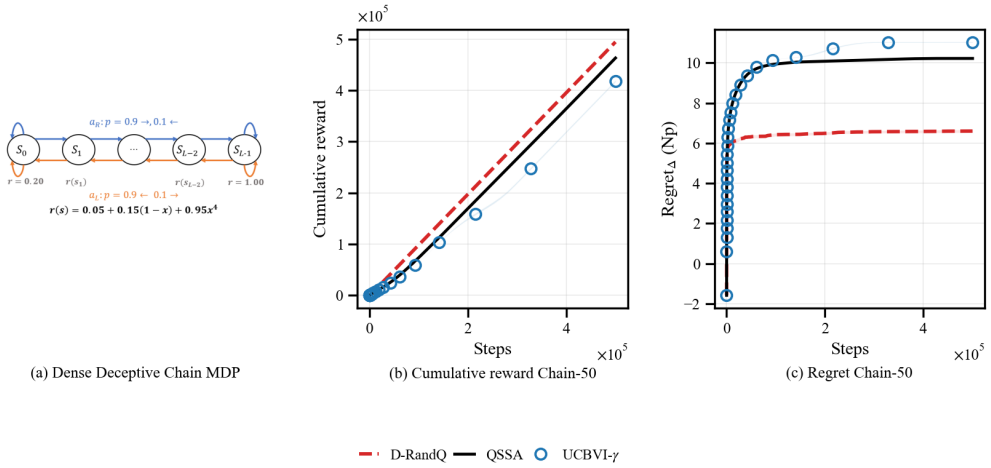
Figure 1: Maze-25: representative successful trajectories, mean cumulative normalized reward, and log mean cumulative gap. Shading: pointwise 95% paired percentile-bootstrap intervals over 30 base seeds (10,000 resamples). No theorem curve is plotted because the regret bound contains generic universal constants that are not numerically instantiated, and the displayed practical parameters are not theorem-certified.

exact reproduction of the schedule analyzed by He et al. (2020) (Section I). All runs use $\gamma = 0.99$, $\delta = 0.05$, and practical settings $J = 16$, $c_b = 0.003$, $c_{\mathrm{VI}} = 0.01$. These multipliers are practical scaling choices rather than theorem-specified constants, selected on a disjoint calibration task under a matched tuning budget and then transferred to Chain-50, so the comparison is not tuning-independent (Section I.5). Curves average 30 paired base seeds; pointwise 95% intervals use 10,000 percentile-bootstrap resamples with jointly resampled seed indices, and for Chain the two recorded action-label permutations are averaged within each base seed before resampling. Seed-0 maze trajectories are qualitative only, and optimal values are computed offline for evaluation. All cumulative action gaps below are recomputed under the common criterion (1) on our instances; these empirical comparisons do not compare the distinct theoretical regret guarantees of the baseline papers. Figures 1–2 show the logarithm of mean cumulative gap, with interval endpoints transformed likewise.

Environments. Maze-25 is the fixed 25×25 wall map with start $(0, 0)$ and goal $(24, 24)$. An intended move occurs with probability 0.9 and each perpendicular move with probability 0.05; walls and boundaries self-loop. Raw rewards $-0.1/0/1$ for self-loop/legal move/goal entry are mapped by $(r + 0.1)/1.1$ to $[0, 1]$, and the goal deterministically returns to the start on the next interaction; we run 2×10^6 continuing interactions. Chain-50 has states $0{:}49$ and starts at 0, never terminating. Left/right moves succeed with probability 0.9 and reverse with probability 0.1; boundaries self-loop, and at pre-transition state s both actions receive $r(s) = 0.05 + 0.15(1 - x) + 0.95x^4$ with $x = s/49$; we run 5×10^5 interactions. Exact wall, seed, baseline, and tuning specifications are in Appendix I.

Maze results. On the maze instance, D-RandQ attains the highest cumulative reward and lowest cumulative gap (Figure 1). This is the primary baseline comparison, because the practical multipliers were selected on the disjoint C-Maze-15 calibration task. Final gaps, ordered as D-RandQ/QSSA/UCBVI-γ, are $(5.92, 9.80, 8.77) \times 10^4$: D-RandQ reduces the gap relative to the two baselines by 39.6% and 32.6%, respectively, and achieves a lower gap in all 30 paired seeds, with cumulative reward 1.22×10^5. The paired endpoint gap differences (D-RandQ minus baseline) are -3.887×10^4 [95% CI $-4.065, -3.725]\times 10^4$ versus QSSA and -2.856×10^4 $[-3.036, -2.690]\times 10^4$ versus UCBVI-γ.

Chain results. At the transferred practical operating point, the separation in the reported Chain-50 runs is larger (Figure 2): final cumulative rewards are $(4.95, 4.63, 4.18) \times 10^5$ and gaps are $(0.074, 2.75, 6.18) \times 10^4$ in the same method order, and D-RandQ's nearly flat late-stage gap curve is consistent with escape from the deceptive left endpoint. Its paired gap differences are -2.673×10^4 [95% CI $-2.692, -2.655]\times 10^4$ versus QSSA and -6.105×10^4 $[-6.188, -5.970]\times 10^4$ versus



(a) Dense Deceptive Chain MDP (b) Cumulative reward Chain-50 (c) Regret Chain-50

Figure 2: Chain-50: means over 30 paired seeds after within-seed action-permutation averaging. Shading: pointwise 95% paired percentile-bootstrap intervals from 10,000 base-seed resamples, transformed with the endpoints in the log-gap panel. No theorem curve is plotted.

UCBVI-γ, with all 30 base-seed differences favoring D-RandQ, but these runs reuse the multipliers selected on the disjoint maze calibration task, and separate dense-chain artifacts show substantial sensitivity of the optimistic baselines to bonus scaling, so the comparison characterizes the reported transferred operating point rather than tuning-independent dominance. We therefore view Chain-50 as an empirical stress test, with Maze-25 providing the cleaner baseline validation.

Mechanism ablations. Under matched practical parameters, D-RandQ has lower cumulative action-gap and higher cumulative reward than Slow-only in all 30 paired base seeds on both instances (mean endpoint gaps $59{,}170.1/83{,}223.1$ on Maze and $736.5/179{,}623.4$ on Chain-50; the learning curves and the optimistic-initialization ablation are in Figures 3–4). This comparison is architectural rather than single-mechanism: Slow-only jointly removes the fast ensemble heads and the fast–slow mixture, replaces the per-step policy refresh by updates of the committed slow table $\widetilde{Q}^\flat$, and replaces the fast bootstrap source $\widehat{V}$ by the committed slow value $\widetilde{V}^\flat$, with target freezing common to both variants, so it does not isolate which component is individually responsible for the improvement. Replacing the Beta draws by their conditional means produced no clearly resolved endpoint difference at these budgets, the paired reward- and gap-difference intervals including zero on both tasks; the Beta randomization remains essential to the present proof, which uses Dirichlet anti-concentration over independently drawn weights. The shared $U = 100, J = 16$ are not theorem-certified parameters, which are $U = 2H_\gamma$ together with a logarithmically growing ensemble width $J = \lceil c_J \ell \rceil$ and the slow-prior parameters of (5): the experiments neither instantiate the certified algorithm numerically nor verify the finite-sample upper bound, and the certified constants are worst-case and conservative (Appendix I.4). This is a gap in constant choices rather than a contradiction of Theorem 1, with a cleaner mechanism isolation left to future work.

6 Conclusion

We studied Discounted Randomized Q-Learning (D-RandQ), which extends RandomizedQ's fast/slow, per-step design to continuing discounted interaction. Our analysis lifts prior stage-wise reuse counting to global interaction-time occurrences, aligning value versions and charging self-loops and boundaries separately, and yields a finite-sample, gap-independent action-gap bound that is sub-linear in T for fixed problem parameters, plus an exact Bellman-telescoping conversion to sublinear Liu–Su γ-regret. The bound also carries an $O(\bar{\eta}T)$ fast–slow mixture charge from the uniform $\bar{\eta}H_\gamma$ shift, and the present analysis does not capture a theoretical agility benefit from the fast path, whose bootstrap contribution is bounded by its global range; sharper fast-path analysis remains future work.

REFERENCES

S. Agrawal and R. Jia. Optimistic posterior sampling for reinforcement learning: Worst-case regret bounds. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1184–1194, 2017.

C. Dann, M. Mohri, T. Zhang, and J. Zimmert. A provably efficient model-free posterior sampling method for episodic reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 12040–12051, 2021.

M. Fortunato, M. G. Azar, B. Piot, J. Menick, I. Osband, A. Graves, V. Mnih, R. Munos, D. Hassabis, O. Pietquin, C. Blundell, and S. Legg. Noisy networks for exploration. In *International Conference on Learning Representations*, 2018.

J. He, D. Zhou, and Q. Gu. Nearly minimax optimal reinforcement learning for discounted MDPs. *arXiv:2010.00587*, 2020.

X. Ji and G. Li. Regret-optimal model-free reinforcement learning for discounted MDPs with short burn-in time. In *Advances in Neural Information Processing Systems*, volume 36, pp. 80674–80689, 2023.

C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, volume 31, 2018.

G. Li, L. Shi, Y. Chen, Y. Gu, and Y. Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17762–17776, 2021.

S. Liu and H. Su. Regret bounds for discounted MDPs. *arXiv:2002.05138*, 2020.

I. Osband, D. Russo, and B. Van Roy. (More) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, volume 26, pp. 3003–3011, 2013.

I. Osband, C. Blundell, A. Pritzel, and B. Van Roy. Deep exploration via bootstrapped DQN. In *Advances in Neural Information Processing Systems*, volume 29, pp. 4026–4034, 2016.

I. Osband and B. Van Roy. Posterior sampling for reinforcement learning without episodes. *arXiv:1608.02731*, 2016.

M. Simchowitz and K. G. Jamieson. Non-asymptotic gap-dependent regret bounds for tabular MDPs. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

R. Singh, S. Chandak, E. Moulines, V. S. Borkar, and N. Bambos. Regret and sample complexity of online Q-learning via concentration of stochastic approximation with time-inhomogeneous Markov chains. *arXiv:2602.16274v2*, 2026.

A. L. Strehl, L. Li, E. Wiewiora, J. Langford, and M. L. Littman. PAC model-free reinforcement learning. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 881–888, 2006.

R. S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1):9–44, 1988.

D. Tiapkin, D. Belomestny, D. Calandriello, E. Moulines, R. Munos, A. Naumov, P. Perrault, M. Valko, and P. Ménard. Model-free posterior sampling via learning rate randomization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 73719–73774, 2023. Revised arXiv version 2310.18186v2, 2025.

H. Wang, X. Xu, and Y. Chi. Provably efficient and agile randomized Q-learning. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 2377–2385, 2026. arXiv:2506.24005v2.

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

C.-Y. Wei, M. Jafarnia-Jahromi, H. Luo, H. Sharma, and R. Jain. Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 10170–10180, 2020.

K. Yang, L. F. Yang, and S. S. Du. Q-learning with logarithmic regret. In *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1576–1584, 2021.

G. Zamir, M. Zurek, and Y. Chen. Optimal variance-dependent regret bounds for infinite-horizon MDPs. In *Proceedings of the Thirty-Ninth Conference on Learning Theory*, volume 336 of *Proceedings of Machine Learning Research*, pp. 7016–7061, 2026.

A. Zanette, D. Brandfonbrener, E. Brunskill, M. Pirotta, and A. Lazaric. Frequentist regret bounds for randomized least-squares value iteration. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1954–1964, 2020.

Z. Zhang, Y. Zhou, and X. Ji. Almost optimal model-free reinforcement learning via reference-advantage decomposition. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

A NOTATION, TECHNICAL CONVENTIONS, AND PARAMETER CHOICES

A.1 NOTATION

This appendix summarizes the notation used in the algorithm, results, and proofs. Bold uppercase and lowercase symbols denote matrices and vectors, respectively; their scalar entries are written in ordinary type. Thus $\boldsymbol{A}$, $\boldsymbol{d}$, $\boldsymbol{r}$, and $\boldsymbol{E}^\flat$ denote matrices or vectors, whereas A_{tj}, d_t, r_t, and $E_t^\flat$ are scalars. A superscript $\flat$ marks the slow ensemble, a hat marks a temporary ensemble estimate, and a tilde marks a committed slow estimate.

Symbol	Meaning
MDP, histories, and performance criteria	
$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$	Finite continuing discounted MDP.
$\mathcal{S}, \mathcal{A}$; S, A	State and action sets; their cardinalities.
$x = (s, a)$	A state–action pair.
$P(\cdot \mid s, a)$, $P_x f$	Transition law and the kernel average $\sum_{s'} P(s' \mid s, a) f(s')$.
$r(s, a)$, R_t	Mean reward and observed reward at transition t; $R_t \in [0, 1]$.
K_x	Joint conditional kernel of the reward and next state for pair x.
γ, H_γ	Discount factor and effective horizon $H_\gamma = (1 - \gamma)^{-1}$.
s_t, a_t	State and selected action at time t.
H_t, $\mathcal{F}_t$	Complete pre-action history and its generated sigma-field.
$\mathbb{E}$, $\mathbb{P}$	Expectation and probability.
Q^*, V^*	Optimal action-value and value functions.
T	Number of analyzed transitions.
g_t, $\mathrm{Regret}_\Delta(T)$	One-step optimal-action gap and its cumulative regret through time T.
$\mathrm{Regret}_\gamma^{\mathrm{LS}}(T)$	Reward-based γ-regret of Liu & Su (2020), defined in (6).
$N_\epsilon(T)$	Number of interactions through T with $g_t \geq \epsilon$, for a fixed threshold $\epsilon > 0$.
Algorithm and parameter choices	
U	Optimistic initialization, equal to $2H_\gamma$.
L	Stage-growth parameter defined in (2).
M_q	Length of slow stage q.
$q(s, a)$	Current slow-stage index of pair (s, a).
$N(s, a)$, $N^\flat(s, a)$	Cumulative fast visit counter and within-stage slow counter.
J, j	Number of ensemble heads and a head index.
$\widehat{Q}^j$, $\widehat{Q}^{\flat,j}$	Fast head and temporary slow head.
$\widetilde{Q}^\flat$, Q	Committed slow table and mixed table used for action selection.
$\widehat{V}$, $\widetilde{V}^\flat$	Fast policy value and committed slow value.
Y_t, $Y_t^\flat$	Frozen fast and slow TD targets at transition t.
w_m^j, $w_m^{\flat,j}$	Fresh beta learning rates for a fast update and a slow update.
η_q, $\eta_t(s, a)$	Stage mixing rate and the rate stored in a policy entry.
$\bar{\eta}$	Uniform upper bound $(L + 1)^{-1/2}$ on all mixing rates.
δ, ℓ	Confidence level and logarithmic factor $\ell = \log(128SAT/\delta)$.
κ, n_0	Fast beta scale and prior parameter; here both equal one.
$\kappa^\flat, n_0^\flat$	Slow beta scale and slow prior parameter.
$c_J, c_\flat, c_0$	Fixed numerical constants used in the parameter choices; $c_\flat = 8$, $c_J = 2/p_0$, and $c_0 \approx 2781.9041$ are explicit.
Dependency recursion and confidence envelopes	
D_t	Selected-action estimation error $Q_t(s_t, a_t) - Q^*(s_t, a_t)$.
d_t, $\boldsymbol{d}$	Shifted scalar error $D_t + \bar{\eta} H_\gamma$ and the vector $(d_1, \ldots, d_T)^\top$.
$E_t^\flat$, $\boldsymbol{E}^\flat$	Slow value error $\widetilde{V}_t^\flat(s_t) - V^*(s_t)$ and its occurrence vector.
r_t, $\boldsymbol{r}$	Nonrecursive scalar charge and the vector of all charges; these are distinct from the reward function $r(s, a)$.
$\boldsymbol{A}$, A_{tj}	Strictly lower-triangular slow occurrence matrix and its scalar entries.
$\mathbf{1}$	All-ones vector in $\mathbb{R}^T$.
$\sigma(i)$	Successor occurrence associated with a regular slow sample; $\sigma(i) = i + 1$.
$N_T(x)$	Number of visits to pair x through time T.
$\mathcal{I}_q(x)$	Visit times belonging to slow stage q of pair x.
b_m, $b_M^\flat$	Fast and slow confidence radii in (32)–(33).

Symbol	Meaning
$p_M^\flat$	Exact slow prior-mass contribution in (34).
$\rho_{\gamma,L}, \Gamma_T$	One-step recursion factor and its finite geometric sum, defined in (1).
$\mathcal{E}_{\mathrm{opt}}$	Event on which every committed slow estimate is optimistic.
$\mathcal{E}_{\mathrm{env}}$	Joint event on which all fast and slow upper envelopes hold.
C	A universal positive constant whose value may change between displays.
Random weights and probabilistic tools	
$\mathrm{Beta}(\alpha,\beta)$, $\mathrm{Dir}(\boldsymbol{\alpha})$	Beta and Dirichlet distributions.
$W_m^r, \bar{W}_m^r$	Scalar random coefficient of coordinate r after m affine updates and its deterministic mean.
$\tau_x(i)$	Time of the $(i+1)$-st visit to pair x.
ξ_i	Centered transition-noise increment at an adaptive visit time.
ξ_t^{LS} in (12)	Centered reward–optimal-value Bellman sample used only in the Liu–Su conversion.
$\mathcal{G}_i, \mathcal{G}_i^+$	Sigma-fields immediately before and after revealing the transition at visit index i.
α_i	Deterministic weight used in the weighted Bernstein inequality.
$\mathcal{K}_{\mathrm{inf}}(\nu,\theta)$	Empirical-likelihood information projection used in the Dirichlet anti-concentration argument.
$\overline{\Phi}(u)$	Standard-normal upper-tail probability $\mathbb{P}\{Z \geq u\}$ for $Z \sim \mathcal{N}(0,1)$; it is non-increasing in u.
$C_0, C_{\mathrm{dir}}, p_0$	Constants in the empirical-likelihood and Dirichlet optimism bounds: $C_0 = 4$, $p_0 = \overline{\Phi}(1)/2$, and the imported $C_{\mathrm{dir}} = c_0^{\mathrm{Tiapkin}}(1/2) \approx 15322.5942$ are explicit; see Appendix A.2.1.

A.2 Technical conventions and parameter choices

Interaction protocol and filtration. All occurrences of $\arg\max$ below use one fixed deterministic ordering of $\mathcal{A}$ to break ties. Thus the selected action is measurable with respect to the current value table. Formally, let $\mathcal{F}_t$ contain the initial state, all observations and all algorithmic random variables revealed strictly before selecting a_t. Thus s_t, the current tables, and the tie-broken greedy action a_t are $\mathcal{F}_t$-measurable. We retain a_t in the conditioning below to identify the selected state–action kernel explicitly. Let H_t denote the complete pre-action history. We set $\mathcal{F}_t = \sigma(H_t)$ up to null sets and interpret conditioning on the history as conditioning on this sigma-field. The transition kernel assumption is

$$\mathbb{P}\big((R_t, s_{t+1}) \in \cdot \mid \mathcal{F}_t, a_t\big) = \mathsf{K}_{s_t,a_t}(\cdot),$$

where the kernel has reward mean $r(s_t,a_t)$ and next-state marginal $P(\cdot \mid s_t, a_t)$. Each fresh beta draw at time t is sampled after the transition, with the prescribed beta law conditional on the pre-draw history, which includes that transition. Conditional on this history, all fresh beta draws made at the same visit are mutually independent across heads and across the fast and slow ensembles; draws made at different visits are generated from independent random seeds. These conventions define all conditional expectations and stopping-time filtrations used below. Linearity of conditional expectation gives, for every bounded f,

$$\mathbb{E}[R_t + \gamma f(s_{t+1}) \mid \mathcal{F}_t, a_t] = r(s_t,a_t) + \gamma P_{s_t,a_t} f.$$

No within-transition independence is used in any Bellman or martingale calculation below.

A.2.1 Universal constants

The constant C_0 in Lemma 6 is certified explicitly as $C_0 = 4$; the verification is in the proof of that lemma. We therefore take the explicit certified choices $c_\flat = \max\{1, 2C_0\} = 8$, c_0 chosen so that $c_0\ell - 1 \geq C_{\mathrm{dir}} + \log_{17/16}(2T/\kappa^\flat)$, and c_J chosen so that $(1-p_0)^{c_J\ell} \leq \delta/(4SAT)$. Indeed, with $x = 128SAT/\delta$, one has $SATe^{-p_0c_J\ell} = SATx^{-p_0c_J} \leq \delta/4$ whenever $p_0c_J \geq 1$; we may take $c_J = 2/p_0$. Here $C_0 = 4$ and $p_0 = \overline{\Phi}(1)/2$ are explicit numerical constants. The one imported constant is the shape threshold of the scalar Dirichlet lemma of Tiapkin et al. (2023, Lemma 10),

whose error parameter enters Tiapkin et al. (2023, Eq. (24)) as

$$c_0(\varepsilon) = \left(\frac{4}{\sqrt{\log(17/16)}} + 8 + \frac{49 \cdot 4\sqrt{6}}{9} \right)^2 \frac{2}{\pi\varepsilon^2} + \log_{17/16}\left(\frac{5}{32\varepsilon^2} \right).$$

Since our imported form of that lemma uses the probability prefactor $1-\varepsilon$ with $\varepsilon = 1/2$, the shape threshold is specialized at the same value, so that $2/(\pi\varepsilon^2) = 8/\pi$ and $5/(32\varepsilon^2) = 5/8$; hence

$$C_{\mathrm{dir}} := c_0^{\mathrm{Tiapkin}}(1/2) = \left(\frac{4}{\sqrt{\log(17/16)}} + 8 + \frac{49 \cdot 4\sqrt{6}}{9} \right)^2 \frac{8}{\pi} + \log_{17/16}\left(\frac{5}{8} \right) \approx 15322.5942.$$

Accordingly the constants C_0, $c_\flat$, c_J, p_0, C_{dir}, and c_0 are all explicit, so every constant needed to instantiate the certified hyperparameters J, $\kappa^\flat$, and $n_0^\flat$ of (5) is available in closed form. All these choices are independent of the MDP and the confidence level except through ℓ. The constants are chosen in order: fix the imported constant C_{dir} and the martingale constant $C_0 = 4$, set $p_0 = \overline{\Phi}(1)/2$, where $\overline{\Phi}(u) = \mathbb{P}\{Z \geq u\}$ for $Z \sim \mathcal{N}(0,1)$, and only afterward choose $c_\flat, c_0, c_J$ by the displayed formulas; in particular c_0 is explicit. Generic constants denoted by C below may depend on these fixed numerical choices, but never on S, A, T, γ, or δ. We stress the resulting distinction: the certified algorithmic hyperparameters J, $\kappa^\flat$, and $n_0^\flat$ of (5) are fully constructive, whereas the constant C of Theorem 1 is a generic multiplicative constant in the displayed regret bound and is not numerically instantiated here. For example, since $T \geq 2$ implies $\ell \geq \log 256$ and $\kappa^\flat \geq 1$, it is enough to take

$$c_0 = 2 + \frac{C_{\mathrm{dir}} + 1}{\log 256} + \frac{1}{\log(17/16)} \approx 2781.9041,$$

Indeed, $\log_{17/16}(2T/\kappa^\flat) \leq \ell/\log(17/16)$, and the displayed choice gives $c_0\ell - 1 \geq C_{\mathrm{dir}} + \ell/\log(17/16)$. Thus the parameter choice is independent of T, S, A, δ except through the displayed occurrence of ℓ in $\kappa^\flat$ and $n_0^\flat$. The decimal $c_0 \approx 2781.9041$ in the display follows from the explicit value $C_{\mathrm{dir}} = c_0^{\mathrm{Tiapkin}}(1/2) \approx 15322.5942$; the exact formulas above, not the displayed decimals, are authoritative. In particular, $c_0\ell > 1$, so $n_0^\flat - \kappa^\flat \geq \kappa^\flat(c_0\ell - 1) > 0$; every denominator $M + n_0^\flat - \kappa^\flat$ used in the prior-atom correction is therefore strictly positive.

B PROOFS OF THE MAIN RESULTS

Proof roadmap. The analysis has optimism and concentration components. For optimism, the stage-counter invariant in Lemma 2, the exact weight representation in Lemma 3, and the conditional head independence in Lemma 7 give the stage-reset representation in Lemma 8. Combining it with the adaptive-visit kernel in Lemma 4 and the Dirichlet anti-concentration result in Lemma 6 yields simultaneous slow optimism in Lemma 9, from which Lemma 10 obtains range control and approximate policy optimism.

On the concentration branch, Lemmas 3, 4, 5, 10, 11, and 12 combine to give the fast and slow envelopes in Lemma 13. The stage arithmetic of Lemma 15 and the slow envelope then produce the occurrence matrix in Lemma 16. Finally, Lemmas 3, 13, 14, 15, and 16 yield the charge recursion in Lemma 17. Theorem 1 intersects the optimism and envelope events and invokes Lemmas 14 and 17 to close the finite matrix recursion. The fixed-discount asymptotics remark then gives the asymptotics.

Proof of Theorem 1. First verify the claimed contraction for the prescribed L. If $0 < \gamma < 1/3$, then $L \geq 1$ gives $\rho_{\gamma,L} \leq 3\gamma < 1$. If $1/3 \leq \gamma < 1/2$, then

$$T\gamma H_\gamma^2 \geq \frac{2\gamma}{(1-\gamma)^2} > 1,$$

so $\log(TH_\gamma^2) > \log(1/\gamma)$ and hence $L \geq 2$. Thus $\rho_{\gamma,L} \leq 2\gamma < 1$. Finally, if $1/2 \leq \gamma < 1$, use $\log(1/\gamma) \leq (1-\gamma)/\gamma$ and $TH_\gamma^2 \geq 8$ to obtain

$$L \geq \frac{\log(TH_\gamma^2)}{\log(1/\gamma)} \geq \frac{\gamma}{1-\gamma}\log(TH_\gamma^2) \geq \frac{\gamma \log 8}{1-\gamma} > \frac{2\gamma}{1-\gamma}.$$

It follows that $\gamma + 2\gamma/L < 1$. These three cases prove $0 < \rho_{\gamma,L} < 1$ throughout the allowed parameter range and justify the geometric-series formula. The proof below uses the exact finite sum.

Let $\mathcal{E}_{\mathrm{opt}}$ be the slow-optimism event of Lemma 9, and let $\mathcal{E}_{\mathrm{env}}$ be the joint envelope event of Lemma 13. By the stated bounds,

$$\mathbb{P}(\mathcal{E}_{\mathrm{opt}}^c) \le \delta/2, \qquad \mathbb{P}(\mathcal{E}_{\mathrm{env}}^c) \le \delta/2.$$

Thus, by the union bound,

$$\mathbb{P}(\mathcal{E}_{\mathrm{opt}} \cap \mathcal{E}_{\mathrm{env}}) \ge 1 - \delta.$$

Write $\mathcal{E} = \mathcal{E}_{\mathrm{opt}} \cap \mathcal{E}_{\mathrm{env}}$. On this event, Lemmas 14 and 17 give $d_t \ge g_t \ge 0$, $E_j^\flat \ge 0$, and $r_t \ge 0$. Hence every vector in the following recursion is coordinatewise nonnegative. Since $\boldsymbol{A}$ is strictly lower triangular, its T-th matrix power is zero. Iterating (45) therefore gives

$$\boldsymbol{d} \le \sum_{k=0}^{T-1} (\gamma \boldsymbol{A})^k (\boldsymbol{r} + 3\bar{\eta} H_\gamma \mathbf{1}).$$

For every nonnegative vector $\boldsymbol{v}$, the induced ℓ_1 norm satisfies

$$\mathbf{1}^\top \boldsymbol{A}^k \boldsymbol{v} = \|\boldsymbol{A}^k \boldsymbol{v}\|_1 \le \|\boldsymbol{A}\|_1^k \|\boldsymbol{v}\|_1 \le \left(1 + \frac{2}{L}\right)^k \mathbf{1}^\top \boldsymbol{v}.$$

Applying this to each term with $\boldsymbol{v} = \boldsymbol{r} + 3\bar{\eta} H_\gamma \mathbf{1}$ yields

$$\begin{aligned} \mathbf{1}^\top \boldsymbol{d} &\le \sum_{k=0}^{T-1} \left[\gamma \left(1 + \frac{2}{L}\right)\right]^k \mathbf{1}^\top (\boldsymbol{r} + 3\bar{\eta} H_\gamma \mathbf{1}) \\ &= \Gamma_T \{\mathbf{1}^\top \boldsymbol{r} + 3\bar{\eta} H_\gamma T\}, \end{aligned}$$

because $\mathbf{1}^\top \mathbf{1} = T$. Lemma 14 gives $\mathrm{Regret}_\Delta(T) \le \mathbf{1}^\top \boldsymbol{d}$. Substituting (46) and (3) proves the second term in (7). Finally, $0 \le g_t \le H_\gamma$ deterministically, so $\mathrm{Regret}_\Delta(T) \le H_\gamma T$; combining the two bounds proves the claim. □

Proof of the fixed-discount asymptotics remark. For fixed γ, $L \to \infty$ as $T \to \infty$. Choose $T_0(\gamma)$ so that $2\gamma/L \le (1-\gamma)/2$. Then $\rho_{\gamma,L} \le (1+\gamma)/2$ and the geometric-series claim follows. Also $L \le 1 + H_\gamma \log(T H_\gamma^2)$ because $\log(1/\gamma) \ge 1 - \gamma$. Substitution in (7) proves the displayed asymptotic statement. More explicitly, for fixed S, A, γ, δ one has $L = \Theta_\gamma(\log T)$ and $\ell = \Theta(\log T)$, so the three terms are respectively

$$O_\gamma\left(\frac{T}{\sqrt{\log T}}\right), \qquad O_\gamma(\sqrt{T}\,\log^3 T), \qquad O_\gamma(\log^4 T),$$

each of which is $o(T)$. The constant in $O_\gamma(\cdot)$ may depend on the fixed discount; no uniform bound in γ is asserted.

For the confidence schedule $\delta_T = 1/T$, define the explicit deterministic deviation

$$B_T(\delta) = 2CH_\gamma^2 \left[\frac{T}{\sqrt{L+1}} + L\{\ell^2 \sqrt{SAT} + SA\ell^3\}\right], \qquad \ell = \log \frac{128SAT}{\delta},$$

where C is the universal constant of Theorem 1; then

$$\mathbb{P}\{\mathrm{Regret}_\Delta(T) \le B_T(\delta)\} \ge 1 - \delta. \tag{9}$$

For this schedule, one still has $\ell_T = \log(128SAT^2) = O(\log T)$ for fixed S, A, and therefore $B_T(\delta_T)/T \to 0$. For any $\epsilon > 0$, $B_T(\delta_T)/T \le \epsilon$ for all sufficiently large T. At such horizons, (9) applied with $\delta = \delta_T$ gives

$$\mathbb{P}\{\mathrm{Regret}_\Delta(T)/T > \epsilon\} \le \mathbb{P}\{\mathrm{Regret}_\Delta(T) > B_T(\delta_T)\} \le \delta_T = 1/T \longrightarrow 0,$$

that is,

$$\frac{\mathrm{Regret}_\Delta(T)}{T} \xrightarrow[T\to\infty]{\mathbb{P}} 0. \tag{10}$$

This is convergence in probability. The argument uses separate good events and separately parameterized algorithms at different horizons; it makes no simultaneous-in-T claim.

Since $g_t \geq 0$ deterministically, $\epsilon \mathbf{1}\{g_t \geq \epsilon\} \leq g_t$; summing gives $\epsilon N_\epsilon(T) \leq \mathrm{Regret}_\Delta(T)$. For fixed $\epsilon > 0$ and any $u > 0$, the same algorithm sequence therefore satisfies

$$\mathbb{P}\{N_\epsilon(T)/T > u\} \leq \mathbb{P}\{\mathrm{Regret}_\Delta(T)/T > \epsilon u\} \longrightarrow 0$$

by (10), proving

$$\frac{N_\epsilon(T)}{T} \xrightarrow[T\to\infty]{\mathbb{P}} 0. \tag{11}$$

□

C Conversion to Liu–Su reward-based γ-regret

The following lemma applies to any action sequence adapted to the interaction protocol of Section 2; it is not specific to either algorithm.

Lemma 1 (Action gaps and Liu–Su γ-regret)**.** *Define*

$$\xi_t^{\mathrm{LS}} = R_t + \gamma V^*(s_{t+1}) - \{r(s_t, a_t) + \gamma P_{s_t,a_t} V^*\}. \tag{12}$$

Then ξ_t^{LS} is $\mathcal{F}_{t+1}$-measurable and satisfies $\mathbb{E}[\xi_t^{\mathrm{LS}} \mid \mathcal{F}_t] = 0$. Equivalently, $(\xi_t^{\mathrm{LS}})_{t\geq 1}$ is a martingale-difference sequence with respect to the shifted filtration $(\mathcal{F}_{t+1})_{t\geq 0}$. Its conditional support has width at most H_γ, and, pathwise,

$$\mathrm{Regret}_\gamma^{\mathrm{LS}}(T) = \mathrm{Regret}_\Delta(T) + \gamma\{V^*(s_{T+1}) - V^*(s_1)\} - \sum_{t=1}^{T} \xi_t^{\mathrm{LS}}. \tag{13}$$

Consequently, for every $\eta \in (0,1)$, with probability at least $1-\eta$,

$$\mathrm{Regret}_\gamma^{\mathrm{LS}}(T) \leq \mathrm{Regret}_\Delta(T) + \gamma H_\gamma + H_\gamma \sqrt{\frac{T}{2} \log \frac{1}{\eta}}. \tag{14}$$

Moreover,

$$\left|\mathbb{E}\,\mathrm{Regret}_\gamma^{\mathrm{LS}}(T) - \mathbb{E}\,\mathrm{Regret}_\Delta(T)\right| \leq \gamma H_\gamma. \tag{15}$$

Proof. Since a_t is $\mathcal{F}_t$-measurable, the conditional-kernel convention gives

$$\mathbb{E}[\xi_t^{\mathrm{LS}} \mid \mathcal{F}_t] = \mathbb{E}[R_t + \gamma V^*(s_{t+1}) \mid \mathcal{F}_t] - r(s_t, a_t) - \gamma P_{s_t,a_t} V^* = 0.$$

Since R_t and s_{t+1} are revealed after a_t is selected, ξ_t^{LS} is $\mathcal{F}_{t+1}$-measurable. Together with the display above, this makes $(\xi_t^{\mathrm{LS}})_{t\geq 1}$ a martingale-difference sequence with respect to the shifted filtration $(\mathcal{F}_{t+1})_{t\geq 0}$. Also $R_t + \gamma V^*(s_{t+1}) \in [0, 1+\gamma H_\gamma] = [0, H_\gamma]$. Conditionally on $\mathcal{F}_t$, subtracting its conditional mean therefore places ξ_t^{LS} in an interval of width at most H_γ.

Rearranging (12) and using the Bellman optimality equation gives

$$\begin{aligned}(1-\gamma)V^*(s_t) - R_t &= V^*(s_t) - r(s_t, a_t) - \gamma P_{s_t,a_t} V^* \\ &\quad + \gamma\{V^*(s_{t+1}) - V^*(s_t)\} - \xi_t^{\mathrm{LS}} \\ &= g_t + \gamma\{V^*(s_{t+1}) - V^*(s_t)\} - \xi_t^{\mathrm{LS}}.\end{aligned}$$

Summing over t proves (13). The boundary term is at most γH_γ. Conditional Hoeffding's lemma for increments whose conditional support has width at most H_γ yields

$$\mathbb{P}\left\{-\sum_{t=1}^{T} \xi_t^{\mathrm{LS}} > H_\gamma \sqrt{\frac{T}{2} \log \frac{1}{\eta}}\right\} \leq \eta,$$

which proves (14). Finally, taking expectations in (13) removes the martingale sum, and $|V^*(s_{T+1}) - V^*(s_1)| \leq H_\gamma$ proves (15). □

Proof of Corollary 1. Run Algorithm 1 with confidence parameter $\delta/2$. Then Theorem 1 applies with $\ell = \log \frac{128SAT}{\delta/2} = \log \frac{256SAT}{\delta} = \ell_{\mathrm{LS}}$, so, using only the second branch of the minimum in (7), with probability at least $1 - \delta/2$,

$$\mathrm{Regret}_{\Delta}(T) \le C\Gamma_T H_\gamma \left[\frac{T}{\sqrt{L+1}} + L\{\ell_{\mathrm{LS}}^2 \sqrt{SAT} + SA\ell_{\mathrm{LS}}^3\} \right].$$

By Lemma 1 with $\eta = \delta/2$, with probability at least $1 - \delta/2$,

$$\mathrm{Regret}_{\gamma}^{\mathrm{LS}}(T) \le \mathrm{Regret}_{\Delta}(T) + \gamma H_\gamma + H_\gamma \sqrt{\frac{T}{2} \log \frac{2}{\delta}}.$$

On the intersection of these two events, which has probability at least $1 - \delta$, combining the two displays gives

$$\mathrm{Regret}_{\gamma}^{\mathrm{LS}}(T) \le C\Gamma_T H_\gamma \left[\frac{T}{\sqrt{L+1}} + L\{\ell_{\mathrm{LS}}^2 \sqrt{SAT} + SA\ell_{\mathrm{LS}}^3\} \right] + \gamma H_\gamma + H_\gamma \sqrt{\frac{T}{2} \log \frac{2}{\delta}}.$$

Independently of any probabilistic event, the pathwise bound $\mathrm{Regret}_{\gamma}^{\mathrm{LS}}(T) \le T$ holds: since $V^*(s_t) \le H_\gamma$ and $R_t \ge 0$, each summand of (6) satisfies $(1-\gamma)V^*(s_t) - R_t \le (1-\gamma)H_\gamma = 1$. Combining this deterministic bound with the high-probability bound above proves (8) with probability at least $1 - \delta$. The displayed asymptotic follows from $L = \Theta_\gamma(\log T)$ and $\ell_{\mathrm{LS}} = \Theta(\log T)$ for fixed S, A, γ, δ, exactly as in the fixed-discount asymptotics remark. □

D Algorithmic invariants

Lemma 2 (Stage-counter invariant)**.** *Write $q_x := q(s,a)$ for $x = (s,a)$. At every action-selection point, for each pair x the stored stage counter q_x identifies a unique current stage and satisfies*

$$0 \le N^\flat(x) < M_{q_x},$$

and the temporary slow heads for x contain exactly the updates collected in the current stage since its last reset. A visit either increments $N^\flat(x)$ while keeping q_x fixed, or, when the counter reaches M_q for the current value $q = q_x$, commits that stage, sets $q_x \leftarrow q + 1$, resets $N^\flat(x)$ to zero, and opens a fresh stage of length M_{q+1}.

Proof. Initially $q_x = 0$ and $N^\flat(x) = 0 < M_0$ for every pair, with all temporary heads equal to the reset value U. Suppose the invariant holds immediately before a visit to x. The algorithm uses the corresponding fixed length M_{q_x} and increments the counter exactly once. If the counter is still below that length, no reset occurs and the invariant is preserved. If it equals the length, the conditional branch commits the current heads, advances the stage index once, resets the slow counter and every temporary slow head, while leaving the cumulative fast counter and fast heads unchanged, and hence establishes $0 < M_{q+1}$ for the next action-selection point, where q is the completed stage index. Other pairs are unchanged, so induction over visits proves the claim. □

E Exact randomized-weight algebra

For independent

$$w_k \sim \mathrm{Beta}(a/\kappa, (b+k)/\kappa), \qquad k = 0, \dots, m-1,$$

consider $Z_{k+1} = (1 - w_k)Z_k + w_k X_k$ and define

$$W_m^0 = \prod_{k=0}^{m-1} (1 - w_k), \qquad W_m^{i+1} = w_i \prod_{k=i+1}^{m-1} (1 - w_k). \tag{16}$$

For $m = 0$, all products are empty, $W_0^0 = 1$, and the data-coordinate sum is empty.

Lemma 3 (Unrolling, means, and column sums)**.** *For $m \geq 0$ (with the preceding empty-product convention),*

$$Z_m = W_m^0 Z_0 + \sum_{i=0}^{m-1} W_m^{i+1} X_i, \qquad \sum_{i=0}^{m} W_m^i = 1, \tag{17}$$

$$\mathbb{E}W_m^0 = \prod_{k=0}^{m-1} \frac{b+k}{a+b+k}, \qquad \mathbb{E}W_m^{i+1} = \frac{a}{a+b+i} \prod_{k=i+1}^{m-1} \frac{b+k}{a+b+k}. \tag{18}$$

If $a = L+1$ and $b = n_0$, then, for $m \geq 1$ and $i \geq 0$ respectively,

$$\max_{1 \leq r \leq m} \mathbb{E}W_m^r \leq \frac{L+1}{L+n_0+m}, \qquad \sum_{m=i+1}^{\infty} \mathbb{E}W_m^{i+1} = 1 + \frac{1}{L}. \tag{19}$$

If $a = 1$, $b = n_0^\flat$, and $\kappa = \kappa^\flat$, then

$$(W_m^0, \ldots, W_m^m) \sim \mathrm{Dir}\left(\frac{n_0^\flat}{\kappa^\flat}, \underbrace{\frac{1}{\kappa^\flat}, \ldots, \frac{1}{\kappa^\flat}}_{m \text{ entries}}\right), \tag{20}$$

so

$$\mathbb{E}W_m^0 = \frac{n_0^\flat}{n_0^\flat + m}, \qquad \mathbb{E}W_m^r = \frac{1}{n_0^\flat + m} \quad (1 \leq r \leq m). \tag{21}$$

Proof. The affine recursion gives (17) by induction. Independence and $\mathbb{E}w_k = a/(a+b+k)$ give (18). When $a = L+1$, the data weights are increasing in their data index because

$$\frac{\mathbb{E}W_m^{i+1}}{\mathbb{E}W_m^{i+2}} = \frac{b+i+1}{a+b+i} \leq 1.$$

The newest weight is $a/(a+b+m-1) = (L+1)/(L+n_0+m)$, proving the first part of (19). For the second part put $c = b+i+1$ and use Euler's beta integral:

$$\begin{aligned}\sum_{m=i+1}^{\infty} \mathbb{E}W_m^{i+1} &= \frac{a}{a+c-1} \sum_{n=0}^{\infty} \frac{(c)_n}{(a+c)_n} \\ &= \frac{a}{a+c-1}\, {}_2F_1(1, c; a+c; 1) = \frac{a}{a-1} = 1 + \frac{1}{L}.\end{aligned}$$

Here $(c)_n = \Gamma(c+n)/\Gamma(c)$ is the rising factorial. Gauss' formula ${}_2F_1(1, c; a+c; 1) = \Gamma(a+c)\Gamma(a-1)/[\Gamma(a)\Gamma(a+c-1)] = (a+c-1)/(a-1)$ gives the last equality after cancellation. Gauss' identity is applicable because $a > 1$. Finally, the gamma representation $G_0 \sim \Gamma(n_0^\flat/\kappa^\flat, 1)$ and $G_r \sim \Gamma(1/\kappa^\flat, 1)$, followed by normalization, has the same successive beta ratios as (16); this proves (20)–(21). Indeed, with $S_r = G_0 + \cdots + G_r$, the ratio G_r/S_r has law $\mathrm{Beta}(1/\kappa^\flat, (n_0^\flat + r - 1)/\kappa^\flat)$ and is independent of the normalized previous coordinates. Thus the stick-breaking ratios coincide with the beta variables in the slow recursion, and their products are exactly the Dirichlet coordinates in (20). □

F Probabilistic inputs

Lemma 4 (Kernel preservation at adaptive visit times)**.** *Fix a state–action pair $x = (s, a)$ and let τ_i be the time of its i-th visit, with $\tau_i = \infty$ if that visit does not occur. On $\{\tau_i < \infty\}$, conditional on the pre-transition sigma-field $\mathcal{F}_{\tau_i}$, the pair $(R_{\tau_i}, s_{\tau_i+1})$ has law K_x, independently of how the stopping time τ_i was selected. Consequently, for every bounded measurable f,*

$$\mathbb{E}[f(R_{\tau_i}, s_{\tau_i+1}) \mid \mathcal{F}_{\tau_i}, \tau_i < \infty] = \int f \, d\mathsf{K}_x.$$

Proof. The event $\{\tau_i = t\}$ is measurable with respect to the history before the transition at time t, because the action a_t is selected before the transition is sampled. On this event, the MDP kernel assumption gives $\mathcal{L}((R_t, s_{t+1}) \mid \mathcal{F}_t, a_t) = \mathsf{K}_x$. Summing the conditional identity over t proves the formula and establishes the martingale-difference property for sequentially revealed samples. □

The next result requires one-step predictability: each target may depend on earlier learning rates but is frozen before the current rate is sampled.

Lemma 5 (Predictable beta stochastic approximation)**.** *Let $T \geq 1$ be an integer, $\varepsilon \in (0,1)$, and let $a, b, \kappa \geq 1$. Assume $a + b + i \geq 2\kappa$ for every index used (for the fast parameters this is $L + 2 + i \geq 2$, and for the slow parameters it follows from $n_0^\flat/\kappa^\flat \geq c_0\ell > 2$). Let $(\mathcal{F}_i)$ be a filtration, let $0 \leq X_i, Z_0 \leq B$, and suppose X_i and the recursively generated Z_i are $\mathcal{F}_i$-measurable. Assume that the conditional law of w_i given $\mathcal{F}_i$ is exactly $\mathrm{Beta}(a/\kappa, (b+i)/\kappa)$ and that w_i is $\mathcal{F}_{i+1}$-measurable; in particular, the fresh draw is independent of every $\mathcal{F}_i$-measurable target and iterate. Set*

$$Z_{i+1} = (1 - w_i)Z_i + w_i X_i, \qquad W_m^0 = \prod_{k=0}^{m-1}(1 - w_k), \quad W_m^{i+1} = w_i \prod_{k=i+1}^{m-1}(1 - w_k),$$

where $0 \leq i < m \leq T$, empty products equal one, and Z_{i+1} is $\mathcal{F}_{i+1}$-measurable. Let $\bar{W}_m^r = \mathbb{E}W_m^r$; the fixed conditional law of each fresh draw implies that these means are the deterministic beta-product means in (18)*. For $\lambda = \log(8T/\varepsilon)$, simultaneously for $0 \leq m \leq T$, with probability at least $1 - \varepsilon$,*

$$\left| Z_m - \bar{W}_m^0 Z_0 - \sum_{i=0}^{m-1} \bar{W}_m^{i+1} X_i \right| \leq CB \left\{ \sqrt{\frac{\kappa a \lambda}{a + b + m}} + \frac{\kappa a \lambda}{a + b + m} \right\}. \tag{22}$$

Proof. For $m = 0$ the displayed deviation is identically zero, so assume $m \geq 1$. Let $\bar{w}_i = a/(a+b+i)$ and let $\bar{Z}_i$ follow the recursion with $\bar{w}_i$ in place of w_i. Direct subtraction gives

$$Z_{i+1} - \bar{Z}_{i+1} = (1 - \bar{w}_i)(Z_i - \bar{Z}_i) + (w_i - \bar{w}_i)(X_i - Z_i).$$

Induction in the deterministic recursion gives

$$\bar{Z}_m = \bar{W}_m^0 Z_0 + \sum_{i=0}^{m-1} \bar{W}_m^{i+1} X_i,$$

because successive conditioning under the fixed conditional beta laws gives $\bar{W}_m^0 = \prod_{k<m}(1 - \bar{w}_k)$ and $\bar{W}_m^{i+1} = \bar{w}_i \prod_{k=i+1}^{m-1}(1 - \bar{w}_k)$. The comparison iterate $\bar{Z}_m$ uses the realized adaptive targets $X_0, \ldots, X_{m-1}$ and deterministic mean learning rates. It is therefore random and generally differs from the unconditional expectation of Z_m. The formulas for $\bar{W}_m^r$ follow from the tower property. Both recursions are evaluated on the same target sequence, so the subtraction is a pathwise identity and requires no conditioning on future targets. The last term is a martingale difference. Iteration gives deterministic backward coefficients $c_{i,m} = \prod_{k=i+1}^{m-1}(1 - \bar{w}_k)$. More explicitly, for $m \geq 1$,

$$Z_m - \bar{Z}_m = \sum_{i=0}^{m-1} c_{i,m}(w_i - \bar{w}_i)(X_i - Z_i),$$

where $c_{m-1,m} = 1$. Since $X_i, Z_i \in [0, B]$, each predictable multiplier $(X_i - Z_i)$ is bounded by B; hence the conditional MGF bound can be applied successively without conditioning on future targets.

We next derive the beta MGF bound. If $X \sim \mathrm{Beta}(\alpha, \beta)$, $n = \alpha + \beta$, and $\mu = \alpha/n$, integration by parts in the beta density gives the Stein identity

$$\mathbb{E}[(nX - \alpha)f(X)] = \mathbb{E}[X(1 - X)f'(X)].$$

Indeed, for the beta density $p(x)$ on $(0, 1)$,

$$\frac{d}{dx}\{x(1 - x)p(x)\} = (\alpha - nx)p(x).$$

Multiplying by an absolutely continuous f and integrating by parts gives the displayed identity; the boundary term vanishes because $x(1 - x)p(x)$ tends to zero at both endpoints for $\alpha, \beta > 0$. Apply it

to $f(x) = \exp\{\theta(x-\mu)\}$. Since $X(1-X) \le X$, solving the resulting differential inequality gives, for $0 \le \theta < n$,

$$\log \mathbb{E}e^{\theta(X-\mu)} \le \frac{\mu\theta^2}{2(n-\theta)}, \qquad \log \mathbb{E}e^{-\theta(X-\mu)} \le \frac{\mu\theta^2}{2n}. \tag{23}$$

Indeed, writing $M(\theta) = \mathbb{E}e^{\theta(X-\mu)}$, the Stein identity and $X(1-X) \le X$ give

$$nM'(\theta) \le \theta\mathbb{E}[Xe^{\theta(X-\mu)}] = \theta M'(\theta) + \mu\theta M(\theta),$$

hence $(n-\theta)M'(\theta) \le \mu\theta M(\theta)$. Integration and $n-u \ge n-\theta$ for $0 \le u \le \theta$ prove the first inequality. For $M_-(\theta) = \mathbb{E}e^{-\theta(X-\mu)}$, the same identity and $X(1-X) \le X$ give

$$nM_-'(\theta) \le \theta\{\mu M_-(\theta) - M_-'(\theta)\}, \qquad (n+\theta)M_-'(\theta) \le \theta\mu M_-(\theta).$$

Thus $(\log M_-)'(\theta) \le \mu\theta/(n+\theta) \le \mu\theta/n$, and integration proves the second inequality in (23). Consequently $w_i - \bar{w}_i$, conditionally on $\mathcal{F}_i$, is sub-gamma on the positive side with variance factor

$$v_i = \frac{a\kappa}{(a+b+i)^2} \quad \text{and scale} \quad c_i = \frac{\kappa}{a+b+i};$$

the negative-side bound is stronger and obeys the same sub-gamma envelope. This follows by substituting $n = (a+b+i)/\kappa$ and $\mu = a/(a+b+i)$ into (23). The scale term is retained to obtain a uniform bound with this variance factor. Bounded beta variables also admit a coarser global Hoeffding sub-Gaussian bound, but boundedness alone does not give a global sub-Gaussian bound with variance factor $a\kappa/(a+b+i)^2$. Because $a \ge 1$,

$$c_{i,m} \le \prod_{k=i+1}^{m-1}\left(1 - \frac{1}{a+b+k}\right) = \frac{a+b+i}{a+b+m-1}.$$

Consequently, if $q_{i,m} = c_{i,m}(X_i - Z_i)$, then $q_{i,m}$ is predictable and $|q_{i,m}| \le Bc_{i,m}$. Define the deterministic envelopes

$$\bar{V}_m = B^2 \sum_{i<m} c_{i,m}^2 \frac{a\kappa}{(a+b+i)^2}, \qquad \bar{C}_m = B \max_{i<m} \frac{\kappa c_{i,m}}{a+b+i}.$$

Using the preceding product bound,

$$\bar{V}_m \le \frac{CB^2\kappa a}{a+b+m}, \qquad \bar{C}_m \le \frac{CB\kappa}{a+b+m}.$$

The first inequality is obtained by substituting $c_{i,m} \le (a+b+i)/(a+b+m-1)$, which makes every summand at most $B^2 a\kappa/(a+b+m-1)^2$; summing m terms gives the stated order. The second follows by cancelling $a+b+i$ in the same product bound. Here we used $m \le a+b+m$ and $a+b+m-1 \ge (a+b+m)/2$. For either sign of the final sum, the positive beta MGF bound is used when the corresponding multiplier has positive sign and the stronger negative bound is used otherwise. Successive conditional expectation therefore gives, for $0 \le \theta < \bar{C}_m^{-1}$,

$$\log \mathbb{E}\exp\left\{\theta \sum_{i<m} q_{i,m}(w_i - \bar{w}_i)\right\} \le \frac{\theta^2 \bar{V}_m}{2(1-\theta\bar{C}_m)}.$$

The deterministic envelopes permit successive application of the conditional MGF bounds. Chernoff's inequality with $\theta = \sqrt{2\lambda/\bar{V}_m}/\{1 + \bar{C}_m\sqrt{2\lambda/\bar{V}_m}\}$ (with the evident limiting interpretation if $\bar{V}_m = 0$) gives

$$\left|\sum_{i<m} q_{i,m}(w_i - \bar{w}_i)\right| \le \sqrt{2\bar{V}_m\lambda} + \bar{C}_m\lambda$$

outside an event of probability $2e^{-\lambda}$. Substituting the preceding bounds on $\bar{V}_m, \bar{C}_m$ yields $CB\{\sqrt{\kappa a\lambda/(a+b+m)} + \kappa\lambda/(a+b+m)\}$. Since $a \ge 1$, this is no larger than the right-hand side of (22). A union bound over $m \le T$ and the two tails proves the simultaneous statement. The argument uses only predictability at each update, allowing the targets to depend on earlier draws. □

For $b > 0$, let $\mathcal{P}([0, b])$ denote the Borel probability measures on $[0, b]$. We use the extended relative entropy convention

$$\mathrm{KL}(\nu\|\nu') = \begin{cases} \displaystyle\int \log\left(\frac{d\nu}{d\nu'}\right) d\nu, & \nu \ll \nu', \\ +\infty, & \text{otherwise.} \end{cases}$$

For a probability law ν supported on $[0, b]$ and $\mu \in [0, b]$, define the primal quantity

$$\mathcal{K}_{\inf}(\nu, \mu) = \inf_{\substack{\nu' \in \mathcal{P}([0,b]) \\ \mathbb{E}_{\nu'} X \geq \mu}} \mathrm{KL}(\nu\|\nu'). \tag{24}$$

At the endpoints this gives $\mathcal{K}_{\inf}(\nu, 0) = 0$ for every ν, while $\mathcal{K}_{\inf}(\nu, b) = \mathrm{KL}(\nu\|\delta_b)$ because the only law on $[0, b]$ with mean at least b is δ_b. For $\mu \in (0, b)$, the primal quantity has the dual representation

$$\mathcal{K}_{\inf}(\nu, \mu) = \max_{0 \leq \lambda \leq 1/(b-\mu)} \mathbb{E}_\nu \log\{1 - \lambda(X - \mu)\}. \tag{25}$$

The equality is the one-dimensional convex dual of the mean constraint. To see the direction needed below, fix an admissible λ and define $h_\lambda(x) = 1 - \lambda(x - \mu) \geq 0$. For every $\nu' \in \mathcal{P}([0, b])$ with $\mathbb{E}_{\nu'} X \geq \mu$, one has $\mathbb{E}_{\nu'} h_\lambda(X) \leq 1$. The Donsker–Varadhan variational inequality therefore gives

$$\mathrm{KL}(\nu\|\nu') \geq \mathbb{E}_\nu \log h_\lambda(X) - \log \mathbb{E}_{\nu'} h_\lambda(X) \geq \mathbb{E}_\nu \log h_\lambda(X).$$

When h_λ vanishes at b, apply the variational inequality first to $h_\lambda + \varepsilon > 0$. Since $\mathbb{E}_{\nu'}(h_\lambda + \varepsilon) \leq 1 + \varepsilon$, this gives

$$\mathrm{KL}(\nu\|\nu') \geq \mathbb{E}_\nu \log(h_\lambda + \varepsilon) - \log(1 + \varepsilon),$$

and then let $\varepsilon \downarrow 0$. If $\nu(\{b\}) > 0$, the resulting right-hand side is $-\infty$ and the inequality is automatic. This also covers a vanishing factor at the endpoint. Taking the infimum over ν' and then the supremum over λ yields the lower bound in (25). For the reverse direction, whenever the maximizer is interior and $h_\lambda(x) > 0$, define the Gibbs tilt

$$\frac{d\nu_\lambda}{d\nu}(x) = \frac{h_\lambda(x)^{-1}}{\mathbb{E}_\nu[h_\lambda(X)^{-1}]}.$$

The first-order condition for the dual objective is $\mathbb{E}_\nu[(X - \mu)/h_\lambda(X)] = 0$, which is equivalent to $\mathbb{E}_{\nu_\lambda} X = \mu$. It also implies $\mathbb{E}_\nu[h_\lambda(X)^{-1}] = 1$ through $h_\lambda^{-1} = 1 + \lambda(X - \mu)h_\lambda^{-1}$, so substituting this tilt into the definition of KL gives equality with $\mathbb{E}_\nu \log h_\lambda(X)$. If the constraint is inactive, $\lambda = 0$ and $\nu' = \nu$ gives equality. It remains to treat a finite optimum at the right boundary $\lambda_B = 1/(b - \mu)$. Finiteness forces $\nu(\{b\}) = 0$, because $h_{\lambda_B}(b) = 0$. Concavity and right-endpoint optimality give the one-sided derivative $D'(\lambda_B-) \geq 0$ for $D(\lambda) = \mathbb{E}_\nu \log h_\lambda(X)$. The derivative formula below follows by dominated convergence on $\{X \leq \mu\}$ and monotone convergence for the magnitude of the negative part on $\{X > \mu\}$. It is finite at such an optimum. Indeed, in $D'(\lambda_B-) = \mathbb{E}_\nu[(\mu - X)/h_{\lambda_B}(X)]$, the positive part comes from $X \leq \mu$ and is bounded by μ, whereas an infinite negative part would give derivative $-\infty$, contradicting right-endpoint optimality. Hence

$$Z := \mathbb{E}_\nu h_{\lambda_B}(X)^{-1} = 1 - \lambda_B D'(\lambda_B-) \leq 1.$$

Define a probability measure by

$$d\nu'_B = h_{\lambda_B}^{-1}\, d\nu + (1 - Z)\delta_b.$$

The added mass is nonnegative, and

$$b - \mathbb{E}_{\nu'_B} X = \int \frac{b - x}{h_{\lambda_B}(x)}\, \nu(dx) = b - \mu;$$

thus $\mathbb{E}_{\nu'_B} X = \mu$. Since $\nu(\{b\}) = 0$, $\mathrm{KL}(\nu\|\nu'_B) = \mathbb{E}_\nu \log h_{\lambda_B}(X)$. This supplies the reverse inequality at the nontrivial boundary and completes the proof of (25), which is used below to control the prior-mass correction.

Lemma 6 (Dirichlet optimism from empirical likelihood)**.** *Let $M \geq 1$ and let $X_1, \ldots, X_M \in [0, H_\gamma]$ be any slow-stage block of baseline observations from one state–action pair, with the same conditional kernel at each visit and mean $\theta \in [0, H_\gamma]$. Equivalently, for the successive visit stopping times τ_i, the conditional law of X_i given the pre-visit history and the visited pair is a fixed law ν, with $\mathbb{E}_\nu X = \theta$.*

Let $\widehat{\nu}_M = M^{-1}\sum_{i=1}^{M}\delta_{X_i}$. There is an explicit universal constant, namely $C_0 = 4$, such that, simultaneously over all sample sizes used by the algorithm,

$$M\mathcal{K}_{\inf}(\widehat{\nu}_M, \theta) \le C_0\ell \tag{26}$$

with probability at least $1-\delta/4$, simultaneously for every stage block created by the algorithm. For any realized block satisfying (26)*, if W is independently drawn conditional on that block with conditional law*

$$\mathcal{L}(W \mid X_{1:M}) = \mathrm{Dir}\left(\frac{n_0^\flat}{\kappa^\flat}, \underbrace{\frac{1}{\kappa^\flat}, \ldots, \frac{1}{\kappa^\flat}}_{M \text{ entries}}\right),$$

then, for the choices $c_\flat = \max\{1, 2C_0\} = 8$ and c_0 specified in Appendix A.2.1, we have

$$\mathbb{P}\left\{2H_\gamma W_0 + \sum_{i=1}^{M} W_i X_i \ge \theta \mid X_{1:M}\right\} \ge p_0 \tag{27}$$

for the explicit value $p_0 = \frac{1}{2}\overline{\Phi}(1) > 0$, where $\overline{\Phi}$ is the standard-normal upper tail.

Proof. The endpoint cases are deterministic: if $\theta = 0$, the weighted target is nonnegative. Moreover, the nonnegative variable X has mean zero, so $X = 0$ almost surely and $\mathcal{K}_{\inf}(\widehat{\nu}_M, 0) = 0$ by (24); thus (26) also holds. If $\theta = H_\gamma$, then the nonnegative random variable $H_\gamma - X$ has expectation zero, so $X = H_\gamma$ almost surely. Hence $\widehat{\nu}_M = \delta_{H_\gamma}$ and $\mathcal{K}_{\inf}(\widehat{\nu}_M, H_\gamma) = 0$ under the endpoint convention, while

$$2H_\gamma W_0 + \sum_{i=1}^{M} W_i X_i = H_\gamma(1 + W_0) \ge H_\gamma.$$

Hence assume for the remainder of this proof that $\theta \in (0, H_\gamma)$.

The variational identity in (25) can be turned into a time-uniform martingale bound as follows. Fix a stage-start stopping time and write $\mathcal{G}_i$ for the history immediately before its i-th visit to the pair and $\mathcal{G}_i^+$ for the sigma-field after revealing X_i. Then $\mathcal{G}_i \subseteq \mathcal{G}_i^+ \subseteq \mathcal{G}_{i+1}$ and X_i is $\mathcal{G}_i^+$-measurable. For every fixed $\lambda \in [0, (H_\gamma - \theta)^{-1}]$, put

$$L_m(\lambda) = \prod_{i=1}^{m}\{1 - \lambda(X_i - \theta)\}.$$

Write

$$\ell_m(\lambda) := \log L_m(\lambda) = \sum_{i=1}^{m}\log\{1 - \lambda(X_i - \theta)\},$$

with the conventions $\log 0 = -\infty$ and $e^{-\infty} = 0$; for $m = 0$, the empty product and sum give $L_0(\lambda) = 1$ and $\ell_0(\lambda) = 0$. The factors are nonnegative and $\mathbb{E}[1 - \lambda(X_i - \theta) \mid \mathcal{G}_i] = 1$, so, with the reveal fields inserted between consecutive pre-visit fields, $\{L_m(\lambda)\}_m$ is a nonnegative martingale (and remains one after restarting at the stopping time). Mix these martingales with the probability measure having density proportional to $[\lambda(1/(H_\gamma - \theta) - \lambda)]^{-1/2}$ and call the mixture L_m. Equivalently, with $B_\theta = (H_\gamma - \theta)^{-1}$, the normalized mixing measure is

$$\pi(d\lambda) = \frac{d\lambda}{\pi\sqrt{\lambda(B_\theta - \lambda)}}, \qquad 0 < \lambda < B_\theta.$$

Define the common filtration by $\mathcal{H}_0 = \mathcal{G}_1$ and $\mathcal{H}_m = \mathcal{G}_m^+$ for $m \ge 1$. Since $\mathcal{H}_m \subseteq \mathcal{G}_{m+1}$, conditional Tonelli and the tower property give

$$\mathbb{E}[L_{m+1} \mid \mathcal{H}_m] = \mathbb{E}[\mathbb{E}[L_{m+1} \mid \mathcal{G}_{m+1}] \mid \mathcal{H}_m] = L_m.$$

Thus the mixture starts at $L_0 = 1$ and is itself a nonnegative martingale; no uncountable union over λ is taken. To handle a stage that receives fewer than its scheduled number of visits before the calendar horizon, lift this visit-indexed process to calendar time. After the stage starts, let $N_{x,q}(t)$ be the number of its owner-pair visits revealed by time t, capped at the scheduled stage length, and set

$\widetilde{L}_t = L_{N_{x,q}(t)}$. At a non-owner transition (and after the cap is reached) this process is unchanged; at an owner transition its multiplier has conditional mean one by the kernel property above. Hence $\{\widetilde{L}_t\}$ is a nonnegative martingale in the post-transition calendar filtration, stopped at the stage end. Ville's inequality at deterministic calendar time T therefore gives

$$\mathbb{P}\left(\max_{m\le N_{x,q}(T)} L_m \ge u\right) \le u^{-1}.$$

The maximum ranges over revealed prefixes. To bound the mixture integral deterministically, make the change of variable $v = \lambda/B_\theta$ and define

$$P_m(v) = \prod_{i=1}^{m}\{1 - B_\theta v(X_i - \theta)\}, \qquad 0 \le v \le 1.$$

Every factor is nonnegative on $[0,1]$, so P_m is a nonnegative polynomial of degree at most m, and $\max_{v\in[0,1]} P_m(v) = \exp\{m\mathcal{K}_{\inf}(\widehat{\nu}_m, \theta)\}$. Markov's polynomial inequality on $[0,1]$ gives

$$\|P_m'\|_\infty \le 2m^2\|P_m\|_\infty.$$

From a maximizer v_m^* choose the direction having at least one half of the unit interval available. On the adjacent one-sided interval of length $1/(4m^2)$, the mean-value theorem then gives $P_m(v) \ge P_m(v_m^*)/2$. Since the arcsine density satisfies $[\pi\sqrt{v(1-v)}]^{-1} \ge 2/\pi$, for every $m \ge 1$ we obtain

$$\int e^{\ell_m(\lambda)}\,\pi(d\lambda) \ge \frac{e^{m\mathcal{K}_{\inf}(\widehat{\nu}_m,\theta)}}{4\pi m^2}.$$

The chosen one-sided direction also covers endpoint maximizers, so no differentiability of ℓ_m at a vanishing factor is required. Therefore

$$m\mathcal{K}_{\inf}(\widehat{\nu}_m, \theta) > \log u + \log(4\pi m^2) \quad \Longrightarrow \quad L_m > u.$$

Taking $u = 4SAT/\delta$ in Ville's inequality, and enlarging the threshold, gives directly (without a separate peeling union over m)

$$\mathbb{P}\left(\exists\, 1 \le M \le N_{x,q}(T) : M\mathcal{K}_{\inf}(\widehat{\nu}_M, \theta) > \log\frac{8SAT}{\delta} + 3\log(e\pi(2M+1))\right) \le \frac{\delta}{4SAT}.$$

By Lemma 4, this calculation is valid at the random visit times because the observations from a fixed pair have the fixed reward–transition kernel. To make the random-block union precise, enumerate all stages by their chronological start times and pad the list with cemetery stages after the algorithm stops. Let ζ_r be the start time of the r-th actual stage, let $I_r = \mathbf{1}\{\zeta_r \le T\}$, and let $\mathcal{F}_{\zeta_r}$ contain the identity of its state–action owner and all history before its first transition. Conditional on $\mathcal{F}_{\zeta_r}$ and $I_r = 1$, the restarted mixture has initial value one, so the preceding Ville bound gives

$$\mathbb{P}(\mathcal{B}_r^c \mid \mathcal{F}_{\zeta_r}, I_r = 1) \le \frac{\delta}{4SAT},$$

where $\mathcal{B}_r$ is the displayed empirical-likelihood inequality for every prefix of that stage, including the observed prefix of an unfinished final stage. Since each transition can start at most one stage, there are at most T actual stages (and $T \le SAT$ because $S, A \ge 1$). Consequently, by conditional expectation,

$$\mathbb{P}\left(\bigcup_{r\le SAT}\{I_r = 1, \mathcal{B}_r^c\}\right) \le \sum_{r\le SAT}\mathbb{E}[I_r\mathbb{P}(\mathcal{B}_r^c \mid \mathcal{F}_{\zeta_r}, I_r = 1)] \le \frac{\delta}{4}.$$

This argument never conditions on a future stage boundary. It remains to dominate the threshold by a multiple of ℓ. First, $\log(8SAT/\delta) \le \ell$ because $8 < 128$. Second, since $M \le T$ and $T \ge 2$,

$$e\pi(2M+1) \le e\pi(2T+1) < 128T \le \frac{128SAT}{\delta},$$

where the middle inequality holds for every $T \ge 1$ and the last uses $S, A \ge 1$ and $\delta < 1$. Hence $\log(e\pi(2M+1)) \le \log(128SAT/\delta) = \ell$, and the threshold inside the event is at most

$$\log\frac{8SAT}{\delta} + 3\log(e\pi(2M+1)) \le \ell + 3\ell = 4\ell.$$

Thus the explicit choice $C_0 = 4$ proves (26) simultaneously for every random stage block.

We next spell out the prior-atom correction in the Dirichlet lower-tail argument. Put

$$\widetilde{\nu} = \frac{n_0^\flat - \kappa^\flat}{M + n_0^\flat - \kappa^\flat}\delta_{2H_\gamma} + \frac{M}{M + n_0^\flat - \kappa^\flat}\widehat{\nu}_M.$$

Using the second representation in (25), the prior term is nonpositive for every admissible λ, and its admissible interval is a subset of the one for $\widehat{\nu}_M$. Therefore

$$\mathcal{K}_{\inf}(\widetilde{\nu}, \theta) \le \frac{M}{M + n_0^\flat - \kappa^\flat}\mathcal{K}_{\inf}(\widehat{\nu}_M, \theta). \tag{28}$$

Indeed, writing $c = (n_0^\flat - \kappa^\flat)/(M + n_0^\flat - \kappa^\flat)$, the objective at any admissible λ is

$$c \log\{1 - \lambda(2H_\gamma - \theta)\} + (1 - c)\,\mathbb{E}_{\widehat{\nu}_M} \log\{1 - \lambda(X - \theta)\}.$$

The first term is nonpositive because $2H_\gamma \ge \theta$; maximizing over the smaller admissible interval and dropping it gives (28). The numerator $n_0^\flat - \kappa^\flat$ is the base prior shape after extracting the extra unit-shape coordinate required by the imported Dirichlet lemma; the extra coordinate is retained as the atom $2H_\gamma$ and is not discarded. In this step the support parameter in (25) changes from $b = H_\gamma$ for $\widehat{\nu}_M$ to $b = 2H_\gamma$ for $\widetilde{\nu}$; the admissible interval therefore becomes $[0, 1/(2H_\gamma - \theta)]$, which is contained in the original interval.

We use the following reduced-prior form of the scalar Dirichlet lemma of Tiapkin et al. (2023, Lemma 10), specialized to the only value needed here, $\varepsilon = 1/2$. Let $\alpha_0, \alpha_1, \dots, \alpha_M > 0$, set $\bar{\alpha} = \alpha_0 + \sum_{i=1}^M \alpha_i$, and let $W \sim \mathrm{Dir}(\alpha_0+1, \alpha_1, \dots, \alpha_M)$. For $f(0) = b_0$ and $0 \le f(i) \le b \le b_0/2$, define $\bar{\nu} = (\alpha_0\delta_{f(0)} + \sum_{i=1}^M \alpha_i\delta_{f(i)})/\bar{\alpha}$. If $\mu \in (0, b)$ and $\alpha_0 \ge C_{\mathrm{dir}} + \log_{17/16}(2(\bar{\alpha} - \alpha_0))$, then

$$\mathbb{P}\{\textstyle\sum_{i=0}^M W_i f(i) \ge \mu\} \ge \frac{1}{2}\overline{\Phi}\Big(\sqrt{2\bar{\alpha}\,\mathcal{K}_{\inf}(\bar{\nu}, \mu)}\Big).$$

This is the notation used in the cited lemma: the Dirichlet parameter has an extra unit shape, while the empirical measure and the normalizing total use the base weights α_i. The extra atom has value $b_0 \ge \mu$ and is handled by the monotone coupling in the cited proof. The lemma is stated for arbitrary positive real shapes; neither α_i nor $\bar{\alpha} - \alpha_0$ is required to be an integer. Thus the value $M/\kappa^\flat$ below is admissible even when it is nonintegral. The restriction $\mu \in (0, b)$ requires the separate endpoint treatment of θ given above. In the revised arXiv version cited in the bibliography, the original conclusion is $(1-\varepsilon)\overline{\Phi}(\sqrt{2\bar{\alpha}\mathcal{K}_{\inf}})$ and its shape condition is $\alpha_0 \ge c_0(\varepsilon) + \log_{17/16}(2(\bar{\alpha} - \alpha_0))$ with $c_0(\varepsilon)$ defined by Tiapkin et al. (2023, Eq. (24)) and reproduced in Appendix A.2.1. Setting $\varepsilon = 1/2$ therefore specializes both ingredients at once: it produces the stated prefactor $1/2$, and it produces the shape threshold $C_{\mathrm{dir}} := c_0^{\mathrm{Tiapkin}}(1/2) \approx 15322.5942$ used below.

Apply this fact with $\alpha_0 = n_0^\flat/\kappa^\flat - 1$, $\alpha_i = 1/\kappa^\flat$, $f(0) = 2H_\gamma$, $f(i) = X_i$, $b_0 = 2H_\gamma$, $b = H_\gamma$, and $\mu = \theta$. Then the base total is $\bar{\alpha} = (n_0^\flat + M - \kappa^\flat)/\kappa^\flat$ and the induced measure is exactly $\widetilde{\nu}$ above. Since a completed stage has $M \ge 1$, $\bar{\alpha} - \alpha_0 = M/\kappa^\flat > 0$, so the logarithm in the imported condition is well-defined. The condition on α_0 is precisely $n_0^\flat/\kappa^\flat - 1 \ge C_{\mathrm{dir}} + \log_{17/16}(2M/\kappa^\flat)$. To make the extra unit-shape coordinate explicit, let

$$G_{0,\mathrm{main}} \sim \Gamma(\alpha_0, 1), \quad G_{0,\mathrm{extra}} \sim \Gamma(1, 1), \quad G_i \sim \Gamma(\alpha_i, 1)\ (1 \le i \le M),$$

all mutually independent, and set $G_0 = G_{0,\mathrm{main}} + G_{0,\mathrm{extra}}$. Normalizing $(G_0, G_1, \dots, G_M)$ gives the original $W \sim \mathrm{Dir}(n_0^\flat/\kappa^\flat, 1/\kappa^\flat, \dots, 1/\kappa^\flat)$, which is the $(\alpha_0+1, \alpha_1, \dots, \alpha_M)$ vector in the imported lemma. In its change-of-measure proof, the extra coordinate is left until the final monotonicity step: the reduced event is written using $G_{0,\mathrm{main}}, G_1, \dots, G_M$ and its common denominator, while the original event additionally contains $G_{0,\mathrm{extra}}$ in the prior coordinate. Since the prior value is $2H_\gamma$ and $2H_\gamma \ge X_i$, adding this independent nonnegative gamma mass can only increase the weighted average. Thus the reduced event is contained in the event for the original W. Let λ^* maximize the variational formula for $\mathcal{K}_{\inf}(\widetilde{\nu}, \theta)$ over $[0, 1/(2H_\gamma - \theta)]$. We separate the inactive case from the nontrivial one.

Inactive constraint. Suppose $\theta \le \mathbb{E}_{\widetilde{\nu}} X$. Then $\nu' = \widetilde{\nu}$ is feasible for the primal definition (24) with $\mathrm{KL}(\widetilde{\nu} \| \widetilde{\nu}) = 0$, so the mean constraint is inactive, $\mathcal{K}_{\inf}(\widetilde{\nu}, \theta) = 0$, and the dual maximizer is $\lambda^* = 0$.

Substituting into the imported Dirichlet lower bound, whose shape hypothesis is unaffected by this case, gives

$$\frac{1}{2}\overline{\Phi}(0) = \frac{1}{4} \geq \frac{1}{2}\overline{\Phi}(1) = p_0,$$

so (27) holds in this case without any exponential tilt.

Nontrivial case. Hence assume $\theta > \mathbb{E}_{\widetilde{\nu}} X$, so that $\lambda^* \neq 0$. The maximizer cannot be the right endpoint: by (5), $n_0^\flat - \kappa^\flat > 0$, so $\widetilde{\nu}$ assigns strictly positive mass to $2H_\gamma$, while the logarithmic factor at that atom is $\log 0 = -\infty$ at the right endpoint. It is therefore interior and $h_i = 1 - \lambda^*(X_i - \theta) > 0$ for every coordinate, including the reduced prior coordinate. The data-coordinate gamma law with shape $1/\kappa^\flat$ and rate 1 is changed to the gamma law with the same shape and rate h_i; equivalently, its density is multiplied by $\exp\{\lambda^*(X_i - \theta)G_i\}$ and renormalized. The same tilt is applied to the reduced prior gamma variable $G_{0,\mathrm{main}}$ with $h_0 = 1 - \lambda^*(2H_\gamma - \theta)$. Under the tilted law, the reduced event is exactly

$$G_{0,\mathrm{main}}(2H_\gamma - \theta) + \sum_{i=1}^{M} G_i(X_i - \theta) \geq 0,$$

because the common denominator $G_{0,\mathrm{main}} + \sum_{i=1}^{M} G_i$ is positive. Its left-hand side is a sum of independent variables under both the reduced original and tilted laws. The data-coordinate normalizing factors have product $\prod_i h_i^{1/\kappa^\flat}$, and the reduced prior contributes $h_0^{\alpha_0}$. The complete normalizing factor is

$$\exp\left\{\frac{M + n_0^\flat - \kappa^\flat}{\kappa^\flat}\mathbb{E}_{\widetilde{\nu}} \log h_{\lambda^*}(X)\right\} = \exp\left\{\frac{M + n_0^\flat - \kappa^\flat}{\kappa^\flat}\mathcal{K}_{\inf}(\widetilde{\nu}, \theta)\right\}.$$

Reversing the change of measure therefore contributes the factor $\exp\{-(M+n_0^\flat-\kappa^\flat)\mathcal{K}_{\inf}(\widetilde{\nu},\theta)/\kappa^\flat\}$. Applying the imported Dirichlet lower bound with the parameters verified above gives

$$\mathbb{P}\left(2H_\gamma W_0 + \sum_{i=1}^{M} W_i X_i \geq \theta \mid X_{1:M}\right) \geq \frac{1}{2}\overline{\Phi}\left(\sqrt{\frac{2(M + n_0^\flat - \kappa^\flat)}{\kappa^\flat}\mathcal{K}_{\inf}(\widetilde{\nu}, \theta)}\right), \tag{29}$$

provided $n_0^\flat/\kappa^\flat - 1 \geq C_{\mathrm{dir}} + \log_{17/16}(2M/\kappa^\flat)$. Here $C_{\mathrm{dir}} := c_0^{\mathrm{Tiapkin}}(1/2) \approx 15322.5942$ denotes the shape-threshold constant of the cited lemma after the same specialization $\varepsilon = 1/2$ that produced the prefactor $1/2$ in (29): both come from that single choice of ε, applied to the prefactor $1 - \varepsilon$ and to the shape threshold $c_0(\varepsilon)$ of Tiapkin et al. (2023, Eq. (24)), whose explicit form is reproduced in Appendix A.2.1. It is distinct from the algorithmic multiplier c_0 in (5). The gamma split, support, and prior correction verify the parameter mapping to Tiapkin et al. (2023, Lemma 10); we invoke that result for the scalar anti-concentration bound and its shape threshold. Combining (28) with (26) shows that the normal argument in (29) is at most one whenever $\kappa^\flat \geq 2C_0\ell$; with the certified values $C_0 = 4$ and $c_\flat = \max\{1, 2C_0\} = 8$ this reads $\kappa^\flat \geq 8\ell$, which is exactly the certified choice $\kappa^\flat = c_\flat \ell$. Finally $n_0^\flat/\kappa^\flat \geq c_0\ell$ implies the displayed shape condition for every $M \leq T$, since

$$\log_{17/16}(2M/\kappa^\flat) \leq \log_{17/16}(2T/\kappa^\flat).$$

Moreover, on (26) and (28), the square of the normal argument is bounded by

$$\frac{2(M + n_0^\flat - \kappa^\flat)}{\kappa^\flat} \cdot \frac{M}{M + n_0^\flat - \kappa^\flat} \cdot \frac{C_0\ell}{M} = \frac{2C_0\ell}{\kappa^\flat} = \frac{8\ell}{8\ell} = 1,$$

where the last step uses $C_0 = 4$ and $\kappa^\flat = c_\flat \ell = 8\ell$. Thus the normal argument in (29) is at most one, and because $\overline{\Phi}$ is non-increasing, the right-hand side of (29) is at least $\frac{1}{2}\overline{\Phi}(1) = p_0$. Hence (27) holds for the explicit value $p_0 = \overline{\Phi}(1)/2$ fixed in Appendix A.2.1. □

Lemma 7 (Monotonicity and conditional head independence)**.** *Fix an initial value z_0 and weights $w_i \in [0, 1]$. The affine recursion $z_{i+1} = (1 - w_i)z_i + w_i y_i$ is coordinatewise nondecreasing in $(z_0, y_0, \ldots, y_{m-1})$. Fix a pair and one of its slow stages. Its* external pre-commit record *consists of the trajectory, frozen targets, fast draws, counters, earlier committed tables, and the raw draws of other stages, through the instant immediately before the current commit is executed. It excludes*

every slow draw of the current stage and every temporary head computed from those draws. It also excludes the current commit output and the policy and value entries synchronized from that output. Conditional on this record, for a stage completed by time T, the J vectors of temporary beta weights are mutually independent and retain their product beta law.

Proof. The first claim follows by induction because each update is a convex combination with nonnegative coefficients. For the second claim, condition first on the history at the stage start, which fixes its owner, index q, and scheduled length M_q. Represent its slow draws by independent random seeds reserved for the indices (i, j), $0 \le i < M_q$, $1 \le j \le J$. These seeds are independent of the stage-start history and of all subsequent transition and other update seeds. This is an equivalent construction of the fresh-draw protocol, not a change to the algorithm.

Before this stage commits, its temporary heads are not read by the action rule, either target, any counter, or any stopping rule. Conditional on the remaining seeds and the stage-start history, changing the reserved seeds therefore leaves the entire external pre-commit record unchanged. This includes the calendar time of completion, whether completion occurs by T, and any commits of other pairs made in the meantime. Equivalently, one can defer revealing all the reserved seeds until just before the commit and then reconstruct the temporary heads from the recorded targets. The distribution of the external record is unchanged by this deferred revelation. Consequently, conditional on the external pre-commit record,

$$\mathcal{L}\big((w_i^{\flat,j})_{i,j} \mid \cdot\big) = \bigotimes_{i=0}^{M_q-1} \bigotimes_{j=1}^{J} \mathrm{Beta}\left(\frac{1}{\kappa^\flat}, \frac{i + n_0^\flat}{\kappa^\flat}\right).$$

The same conditional law holds after removing the initial conditioning by the tower property, because the record includes the stage-start history and the parameters. Selecting the next completed stage chronologically also preserves this factorization: before its commit, the position of that stage in the commit order is determined without reading its reserved seeds. In particular, conditioning on the target block and on completion by T does not reveal these seeds. Conditioning on the current commit output would reveal a function of them and is not part of the asserted record. □

Lemma 8 (Stage reset representation)**.** *For a fixed pair x, at initialization and immediately after every slow commit, its temporary heads equal U and their counters equal zero. During any ensuing stage of length M_q, each temporary head follows the affine recursion started from U with M_q fresh beta weights. Consequently its final weight vector has the Dirichlet law in* (20)*, independently across heads conditional on the frozen stage targets.*

Proof. The initialization assignment gives the claim for stage zero. Thereafter, the reset assignment in Algorithm 1 is performed immediately after each commit and sets every temporary head for the visited pair to U and $N^\flat$ to zero. No update is made to that pair until its next visit, and by Lemma 2 that next visit belongs to the uniquely identified next stage. Its counter selects the beta parameters $(1/\kappa^\flat, (m^\flat + n_0^\flat)/\kappa^\flat)$ in order $m^\flat = 0, \ldots, M_q - 1$. Lemma 3 then gives the Dirichlet representation; independence is the fresh-randomness part of Lemma 7. □

Lemma 9 (Slow ensemble optimism)**.** *For the choices* (5)*, with probability at least $1 - \delta/2$, simultaneously for every $t \in \{1, \ldots, T+1\}$ and $(s, a) \in \mathcal{S} \times \mathcal{A}$, the stored-table snapshots satisfy*

$$\widetilde{Q}_t^\flat(s, a) \ge Q^*(s, a) \quad \textit{and} \quad \widetilde{V}_t^\flat(s) \ge V^*(s). \tag{30}$$

Here the snapshot at $T + 1$ is taken immediately after all updates associated with transition T, before any subsequent action or update. In particular, an entry created by a commit after transition $c \le T$ is optimistic at its first possible read time $c + 1$ and remains so until it is replaced; entries that have not yet been committed retain their optimistic initial value U. This event controls commits made through transition T and makes no claim about the optimism of new entries created after that horizon.

Proof of Lemma 9. Suppose all earlier slow commits are optimistic. Fix a fresh slow stage of $x = (s, a)$ and let $t_1 < \cdots < t_M$ be the calendar times of its M owner-pair transitions (for an unfinished final stage, the observed prefix). Write

$$Y_i^\flat = R_{t_i} + \gamma \widetilde{V}_{t_i}^\flat(s_{t_i+1}), \qquad X_i = R_{t_i} + \gamma V^*(s_{t_i+1}), \qquad 1 \le i \le M,$$

for the frozen slow targets and the baseline observations at those transitions. Pathwise, $Y_i^\flat \geq X_i$; moreover $X_i \in [0, H_\gamma]$. With $\mathcal{G}_i = \mathcal{F}_{t_i}$ the history immediately before the owner-pair transition at calendar time t_i, Lemma 4 and the Bellman equation give, at every adaptive visit,

$$\mathbb{E}[X_i \mid \mathcal{G}_i] = r(x) + \gamma P_x V^* = Q^*(x).$$

The baseline block therefore satisfies the conditional-kernel and mean assumptions of Lemma 6. By Lemma 8, each committed head is a Dirichlet-weighted average of the prior atom U and these targets, with parameters $n_0^\flat/\kappa^\flat, 1/\kappa^\flat, \ldots, 1/\kappa^\flat$. Lemma 6 gives the probability bound for a head whose targets are X_i. Since $Y_i^\flat \geq X_i$ pathwise, Lemma 7 transfers the same bound to the actual targets $Y_i^\flat$. Let $\mathcal{H}_{x,q}$ be the sigma-field generated by the external pre-commit record defined in Lemma 7. In particular, it includes the final transition and all frozen targets, but none of the current stage's slow draws, temporary heads, commit output, or entries synchronized from that output. The block event $\mathcal{B}_{x,q}$ in (26) is $\mathcal{H}_{x,q}$-measurable: the temporary slow heads are not used by the policy or by any committed value until the stage is over. Conditional on $\mathcal{H}_{x,q}$, the second part of Lemma 7 gives independent fresh weights for the J heads. Hence, on $\mathcal{B}_{x,q}$ and assuming all earlier commits are optimistic, the conditional failure probability of this commit is at most $(1-p_0)^J \leq e^{-p_0 J}$.

Enumerate the completed commits chronologically and pad this list to the deterministic length T with cemetery commits. A cemetery commit has block event equal to the sure event, failure event equal to the empty event, and is declared optimistic; these conventions make the induction and its union bound literal also when fewer than T commits occur. For $n = 1, \ldots, T$, let $\mathcal{B}_n$ denote the empirical-likelihood event for the block producing commit n, and let $\mathcal{O}_{n-1}$ denote the event that the first $n-1$ commits are optimistic. For an actual commit, let c_n be its transition time. Let $\mathcal{H}_n$ be the sigma-field generated by the external pre-commit record for the stage producing commit n, stopped immediately before the commit assignment at c_n. Earlier commits and raw temporary draws of other, possibly overlapping stages are included. The current stage's slow draws and temporary heads are excluded, as are its commit output and all table entries computed from that output. This exclusion ensures that the record reveals no function of the current stage's slow draws. The identity, chronological index, and completion time of this stage are measurable in its external record, by the deferred-revelation argument of Lemma 7. No observation after c_n is included. For a cemetery index take $\mathcal{H}_n$ to be the terminal sigma-field. The events $\mathcal{B}_n$ and $\mathcal{O}_{n-1}$ are $\mathcal{H}_n$-measurable. On $\mathcal{O}_{n-1}$, conditioning on $\mathcal{H}_n$ gives

$$\mathbf{1}_{\mathcal{O}_{n-1}} \mathbb{P}(\mathcal{B}_n \cap \{\text{commit } n \text{ is not optimistic}\} \mid \mathcal{H}_n) \leq \mathbf{1}_{\mathcal{O}_{n-1}} \mathbf{1}_{\mathcal{B}_n} e^{-p_0 J}.$$

Let $\mathcal{A}_n = \mathcal{O}_{n-1} \cap \{\text{commit } n \text{ is not optimistic}\}$ be the event that commit n is the first failed commit (and take $\mathcal{A}_n = \varnothing$ for a cemetery index). These events are disjoint, and the tower property applied separately at each n gives

$$\mathbb{P}(\mathcal{B}_n \cap \mathcal{A}_n) \leq e^{-p_0 J}.$$

Since intersecting with all block events can only decrease probability, a union bound over the possible first failure yields

$$\mathbb{P}\left(\bigcap_{n=1}^{T} \mathcal{B}_n \;\cap\; \{\text{some slow commit fails}\}\right) \leq \sum_{n=1}^{T} \mathbb{P}(\mathcal{B}_n \cap \mathcal{A}_n) \leq T e^{-p_0 J} \leq SAT e^{-p_0 J} \leq \delta/4.$$

Lemma 6 and its stopping-time union bound give $\mathbb{P}(\cap_n \mathcal{B}_n)^c \leq \delta/4$. A final union bound therefore gives total slow-optimism failure probability at most $\delta/2$. Initialization is optimistic because $U = 2H_\gamma \geq Q^*$. Induction over the chronological order of commits through transition T proves the first claim, including the snapshot at $T+1$; maximization over actions proves the value claim. More explicitly, a state that has not yet received a slow commit retains $\widetilde{Q}^\flat(s,a) = U$ for every action and $\widetilde{V}^\flat(s) = U$; hence such entries satisfy the same inequalities without any probabilistic argument. At a stage-ending visit the policy table is synchronized with the newly committed entry before the next action is selected, while the newly incremented stage length is used only on the next visit to that pair. Thus the induction applies to exactly the targets used by the algorithm. □

Remark. The optimism argument combines the imported gamma/Dirichlet lower bound (29) with the prior-atom reduction, conditional head independence, and induction over commits. It relies on anti-concentration of the randomized estimates.

G DETERMINISTIC CONSEQUENCES AND CONFIDENCE ENVELOPES

Lemma 10 (Range and approximate policy optimism)**.** *All algorithmic values lie in* $[0, U]$*. On the event* (30)*, for every stored policy entry,*

$$Q_t(s,a) \geq (1-\bar{\eta})Q^*(s,a). \tag{31}$$

Proof. Since $R_t + \gamma U \leq 1 + 2\gamma H_\gamma = U - 1 \leq U$, the range statement follows by induction through convex updates and maxima. Every fast head is nonnegative, every committed slow entry is optimistic, and the synchronization step in Algorithm 1 ensures that each persistent policy entry uses its current committed slow value. Since $\eta_t(s,a) \leq \bar{\eta}$, we have

$$Q_t(s,a) \geq (1-\eta_t(s,a))\widetilde{Q}_t^\flat(s,a) \geq (1-\bar{\eta})Q^*(s,a).$$

Unvisited entries equal U and satisfy the same inequality. □

Let $\tau_x(i)$ be the time of the $(i+1)$-st visit to pair x, indexed from $i = 0$. Define

$$b_m = CH_\gamma \left\{ \sqrt{\frac{L\ell}{L+m}} + \frac{L\ell}{L+m} \right\}, \tag{32}$$

$$b_M^\flat = CH_\gamma \left\{ \sqrt{\frac{\kappa^\flat \ell}{n_0^\flat + M}} + \frac{\kappa^\flat \ell}{n_0^\flat + M} \right\}, \tag{33}$$

$$p_M^\flat = \frac{2H_\gamma n_0^\flat}{n_0^\flat + M}. \tag{34}$$

Lemma 11 (Transition-noise decomposition)**.** *For a fixed pair* $x = (s,a)$ *and its successive visit times* $\tau_x(i)$*, let*

$$\xi_i = R_{\tau_x(i)} - r(x) + \gamma\big(V^*(s_{\tau_x(i)+1}) - P_x V^*\big).$$

Let $\mathcal{G}_i = \mathcal{F}_{\tau_x(i)}$ *be the history immediately before the transition at the* $(i+1)$*-st visit, and let* $\mathcal{G}_i^+$ *be the sigma-field obtained by additionally revealing that visit's reward and next state (and, harmlessly, all update randomness drawn at that visit). Then* $\mathcal{G}_i \subseteq \mathcal{G}_i^+ \subseteq \mathcal{G}_{i+1}$ *whenever the next visit occurs,* ξ_i *is* $\mathcal{G}_i^+$*-measurable, and* $\mathbb{E}[\xi_i \mid \mathcal{G}_i] = 0$*,* $|\xi_i| \leq H_\gamma$*, and* $\mathbb{E}[\xi_i^2 \mid \mathcal{G}_i] \leq H_\gamma^2$*.*

Proof. By Lemma 4, the conditional mean given $\mathcal{G}_i$ is zero because $\mathbb{E}[R_{\tau_x(i)} \mid \mathcal{G}_i] = r(x)$ and $\mathbb{E}[V^*(s_{\tau_x(i)+1}) \mid \mathcal{G}_i] = P_x V^*$; reward and next state need not be independent. Since $R_t \in [0,1]$ and $V^* \in [0, H_\gamma]$, the centered sum is bounded in absolute value by $1 + \gamma H_\gamma = H_\gamma$. The conditional variance bound follows from the same range bound. □

Lemma 12 (Weighted conditional Bernstein)**.** *Let* $(\mathcal{F}_i)$ *be a filtration and let* ξ_i *be* $\mathcal{F}_{i+1}$*-measurable martingale differences with* $\mathbb{E}[\xi_i \mid \mathcal{F}_i] = 0$*,* $|\xi_i| \leq H_\gamma$*, and* $\mathbb{E}[\xi_i^2 \mid \mathcal{F}_i] \leq H_\gamma^2$*. For a finite deterministic index set and deterministic nonnegative weights* α_i *with* $\sum_i \alpha_i \leq 1$*, put* $V = H_\gamma^2 \sum_i \alpha_i^2$ *and* $c = H_\gamma \max_i \alpha_i$*. Then, for every* $u > 0$*,*

$$\mathbb{P}\left\{ \left| \sum_i \alpha_i \xi_i \right| > \sqrt{2Vu} + \frac{cu}{3} \right\} \leq 2e^{-u}.$$

Proof. For an empty index set the sum is zero; define $V = c = 0$ and the assertion is immediate. The same is true if the index set is nonempty but all weights vanish. Hence suppose $c > 0$. For every $|\theta| < 3/c$, the elementary Bernstein MGF inequality for a centered variable bounded by c gives, conditionally on $\mathcal{F}_i$,

$$\mathbb{E}\big[e^{\theta\alpha_i\xi_i} \mid \mathcal{F}_i\big] \leq \exp\left(\frac{\theta^2 \alpha_i^2 H_\gamma^2}{2(1 - |\theta|c/3)} \right).$$

Multiplying these conditional bounds successively (the weights are deterministic) yields

$$\mathbb{E}\exp\left(\theta \sum_i \alpha_i \xi_i \right) \leq \exp\left(\frac{\theta^2 V}{2(1-|\theta|c/3)} \right).$$

For the upper tail, apply Chernoff's bound with

$$\theta_u = \frac{\sqrt{2u/V}}{1 + (c/3)\sqrt{2u/V}} \in (0, 3/c),$$

using the limiting interpretation when $V = 0$. Direct substitution gives

$$\sum_i \alpha_i \xi_i \le \sqrt{2Vu} + \frac{cu}{3}$$

with probability at least $1 - e^{-u}$. Applying the same calculation to $-\xi_i$ and taking a union bound proves the claim. The argument requires only the conditional martingale-difference and MGF properties. □

Lemma 13 (Fast and slow upper envelopes)**.** *There is an event $\mathcal{E}_{\mathrm{env}}$ with probability at least $1 - \delta/2$ on which, simultaneously for all relevant heads, visits, and commits,*

$$\max_j \{\widehat{Q}_m^j(x) - Q^*(x)\} \le \gamma \sum_{i=0}^{m-1} \bar{W}_m^{i+1} \{\widehat{V}_{\tau_x(i)}(s_{\tau_x(i)+1}) - V^*(s_{\tau_x(i)+1})\} + 2H_\gamma \bar{W}_m^0 + b_m, \tag{35}$$

$$\widetilde{Q}^\flat(x) - Q^*(x) \le \frac{\gamma}{n_0^\flat + M} \sum_{i \in \mathcal{I}_q(x)} \{\widetilde{V}_i^\flat(s_{i+1}) - V^*(s_{i+1})\} + p_M^\flat + b_M^\flat. \tag{36}$$

Here the displayed $\widetilde{Q}^\flat(x)$ denotes one fixed stored entry together with its creation commit $c = c(x, q)$. The set $\mathcal{I}_q(x)$ consists of the $M = M_q$ visit times in the unique completed stage whose stage-ending update at time c created that entry. The inequality is attached to this creation node and remains the applicable envelope at every later pre-action snapshot that reads the entry, until the next commit of x replaces it; the indices q, M, and $\mathcal{I}_q(x)$ remain attached to the creation commit. Unfinished stages have no committed entry in this formula. On the event of Lemma 9, the same displayed slow error also satisfies $\widetilde{Q}^\flat(x) - Q^*(x) \ge 0$ for every committed entry.

Proof of Lemma 13. Fix a pair x and use the visit-time filtration from Lemma 4. Write $X_i^x = R_{\tau_x(i)} + \gamma V^*(s_{\tau_x(i)+1})$ and $e_i^x = \widehat{V}_{\tau_x(i)}(s_{\tau_x(i)+1}) - V^*(s_{\tau_x(i)+1})$. Then $X_i^x - Q^*(x) = \xi_i$ with ξ_i as in Lemma 11, and the frozen fast target satisfies $Y_i - Q^*(x) = \xi_i + \gamma e_i^x$. Unroll the corresponding recursion with Lemma 3. For this paragraph, Z_k denotes the scalar recursion $\widehat{Q}_k^j(x)$ for the fixed head and pair. All targets and iterates lie in $[0, U]$ by Lemma 10, so Lemma 5 applies with $B = U = 2H_\gamma$. After m visits,

$$\widehat{Q}_m^j - Q^* = \bar{W}_m^0(\widehat{Q}_0^j - Q^*) + \gamma \sum_{i<m} \bar{W}_m^{i+1} \{\widehat{V}_{\tau_x(i)}(s_{\tau_x(i)+1}) - V^*(s_{\tau_x(i)+1})\}$$
$$+ \underbrace{\sum_{i<m} \bar{W}_m^{i+1} \xi_i}_{T_m^x} + \underbrace{\left[Z_m - \bar{W}_m^0 Z_0 - \sum_{i<m} \bar{W}_m^{i+1} Y_i \right]}_{R_m^j}.$$

Here T_m^x is the transition-noise term and R_m^j is the randomized-rate deviation controlled by Lemma 5. Relabeling the M visits in one slow stage as $i = 1, \dots, M$ with actual times $\tau_1, \dots, \tau_M$, the same decomposition for a slow head after a stage of length M has deterministic data coefficients $(n_0^\flat + M)^{-1}$, prior coefficient $n_0^\flat/(n_0^\flat + M)$, and therefore, writing $e_i^{x,\flat} = \widetilde{V}_{\tau_i}^\flat(s_{\tau_i+1}) - V^*(s_{\tau_i+1})$,

$$\widehat{Q}_M^{\flat,j} - Q^*(x) = \frac{\gamma}{n_0^\flat + M} \sum_{i=1}^M e_i^{x,\flat} + \frac{1}{n_0^\flat + M} \sum_{i=1}^M \xi_i + R_M^{\flat,j} + \frac{n_0^\flat}{n_0^\flat + M}(U - Q^*(x)).$$

Here $R_M^{\flat,j}$ is the slow beta-rate deviation controlled by Lemma 5. This decomposition holds pathwise. The last term is nonnegative and is bounded by $p_M^\flat$ because $0 \le U - Q^*(x) \le 2H_\gamma$. At the i-th visit the variables in the transition sum are precisely $R_i - r(x) + \gamma\{V^*(s_{i+1}) - P_x V^*\}$, after relabeling the visit time as i. If α_i denotes its deterministic mean weight, then $\sum_i \alpha_i \le 1$ (the missing mass

is the initialization coordinate) and $\sum_i \alpha_i^2 \le \max_i \alpha_i$. Bernstein's inequality therefore gives a radius $CH_\gamma\{\sqrt{\lambda \max_i \alpha_i} + \lambda \max_i \alpha_i\}$. For $m = 0$ both transition sums are empty and need no concentration. For fast prefixes with $m \ge 1$, Lemma 3 gives $\max_i \alpha_i \le (L+1)/(L+n_0+m)$; for slow weights it equals $1/(n_0^\flat + M)$. Substituting these bounds gives (32)–(33). The beta-concentration parameter mapping is explicit: for a fast sequence use $(a, b, \kappa, B) = (L+1, 1, 1, 2H_\gamma)$, so $a + b + i = L + 2 + i \ge 2\kappa$ and $a + b + m \ge L + m$; for a reset slow sequence use $(a, b, \kappa, B) = (1, n_0^\flat, \kappa^\flat, 2H_\gamma)$, so $a+b+i \ge 2\kappa^\flat$ follows from $n_0^\flat/\kappa^\flat > 2$, and $a+b+m \ge n_0^\flat + m$. Consequently both the square-root and linear terms in Lemma 5 are covered by (32)–(33) once the logarithms below are bounded by a universal multiple of ℓ.

Concentration for prefixes that may not be reached. We justify applying the fixed-length concentration arguments when the owner pair may cease to be visited. Fix a pair and a fast sequence, or condition on the starting history $\mathcal{R}$ of a particular slow stage. In the latter case $\mathcal{R}$ fixes the owner, stage length, and beta parameters. The starting history is taken before the first transition of the sequence; for a reset it may be the pre-action snapshot immediately after the preceding commit. All probabilities below are conditional on this starting history, on the event that it occurs by T. Fix a prefix length $1 \le m \le T$, also requiring $m \le M_q$ for a slow stage. Let $N_{\mathrm{blk}}(T)$ count the owner visits in this sequence revealed by T, capped at its stage length in the slow case. We will bound failure jointly with $\{N_{\mathrm{blk}}(T) \ge m\}$, without conditioning on this future event.

For each calendar transition t after the start, let I_t indicate that this is an owner visit in the chosen sequence with local index $i_t < m$. Here local indices start at zero. Set $I_t = 0$ after the m-th such visit and after the stage ends. Both I_t and i_t on $\{I_t = 1\}$ are $\mathcal{F}_t$-measurable. Each index $i < m$ is used at most once. Define $\mathcal{D}_t$ to be $\mathcal{F}_t$ augmented by the reward and next state at t, before any learning-rate draw. Thus $\mathcal{F}_t \subseteq \mathcal{D}_t \subseteq \mathcal{F}_{t+1}$, and all targets are $\mathcal{D}_t$-measurable.

First fix deterministic weights $\alpha_0, \dots, \alpha_{m-1}$ for this prefix. At an indicated visit set

$$\Delta_t = \alpha_{i_t}\{R_t - r(x) + \gamma[V^*(s_{t+1}) - P_x V^*]\},$$

and set $\Delta_t = 0$ when $I_t = 0$; no observation at a nonexistent visit is defined. Lemma 4 gives conditional mean zero. Put

$$v_t = I_t H_\gamma^2 \alpha_{i_t}^2, \qquad V = H_\gamma^2 \sum_{i<m} \alpha_i^2, \qquad c = H_\gamma \max_{i<m} \alpha_i,$$

where $v_t = 0$ when $I_t = 0$. We have $\sum_{t \le T} v_t \le V$ pathwise, and the conditional Bernstein MGF calculation gives, for either sign and $0 \le \lambda < 3/c$,

$$\mathbb{E}[e^{\pm\lambda\Delta_t} \mid \mathcal{F}_t] \le \exp\left\{\frac{\lambda^2 v_t}{2(1-\lambda c/3)}\right\}.$$

Consequently the exponentials of the partial sums, compensated by $\lambda^2 \sum_t v_t/[2(1-\lambda c/3)]$, are nonnegative supermartingales starting at one at $\mathcal{R}$. The deterministic upper budget V implies

$$\mathbb{E}\left[e^{\pm\lambda \sum_{t \le T} \Delta_t} \mid \mathcal{R}\right] \le \exp\left\{\frac{\lambda^2 V}{2(1-\lambda c/3)}\right\}.$$

This follows by multiplying the compensated exponential by its compensator, which is bounded above by the displayed deterministic exponential. In particular, it does not replace a random conditional variance by its expectation. On $\{N_{\mathrm{blk}}(T) \ge m\}$, the stopped calendar sum equals the required sum $\sum_{i<m} \alpha_i \xi_i$. Chernoff's argument therefore gives, for every $u > 0$,

$$\mathbb{P}\left\{N_{\mathrm{blk}}(T) \ge m,\ \left|\sum_{i<m} \alpha_i \xi_i\right| > \sqrt{2Vu} + cu/3 \,\middle|\, \mathcal{R}\right\} \le 2e^{-u}. \tag{37}$$

The prefix sum inside this event is evaluated only when the prefix exists. The case $c = 0$ is deterministic and requires no MGF calculation.

For a fixed head, the beta deviation is handled in the same calendar filtration. Fix m and use the deterministic coefficients $c_{i,m}$ from Lemma 5. At an indicated visit let Z_{i_t} be the head before its update and Y_t its frozen target, and set

$$\Delta_t^\beta = c_{i_t,m}(w_{i_t} - \bar{w}_{i_t})(Y_t - Z_{i_t});$$

set it to zero otherwise. With $B = U$, define, on indicated visits,

$$v_t^\beta = B^2 c_{i_t,m}^2 \frac{a\kappa}{(a+b+i_t)^2},$$

and zero otherwise. The budgets from Lemma 5 satisfy

$$\sum_{t\le T} v_t^\beta \le \bar{V}_m = B^2 \sum_{i<m} c_{i,m}^2 \frac{a\kappa}{(a+b+i)^2}, \qquad \bar{C}_m = B \max_{i<m} \frac{\kappa c_{i,m}}{a+b+i}.$$

Conditional on $\mathcal{D}_t$, the multiplier $Y_t - Z_{i_t}$ is fixed and bounded by B in absolute value. Equation (23) gives

$$\mathbb{E}[e^{\pm\lambda\Delta_t^\beta} \mid \mathcal{D}_t] \le \exp\left\{ \frac{\lambda^2 v_t^\beta}{2(1-\lambda\bar{C}_m)} \right\}, \qquad 0 \le \lambda < \bar{C}_m^{-1}.$$

Since v_t^β is already $\mathcal{F}_t$-measurable, the tower property gives the same bound conditional on $\mathcal{F}_t$. The compensated exponential construction just used, now with budgets $\bar{V}_m, \bar{C}_m$, and the pathwise identity in Lemma 5 imply

$$\mathbb{P}\left\{ N_{\mathrm{blk}}(T) \ge m,\ |Z_m - \bar{W}_m^0 Z_0 - \sum_{i<m} \bar{W}_m^{i+1} Y_i| > \sqrt{2\bar{V}_m u} + \bar{C}_m u \,\middle|\, \mathcal{R} \right\} \le 2e^{-u}. \tag{38}$$

Here Y_i denotes the frozen target at local visit i, only on the event that the prefix exists. Applying the deterministic budget estimates of Lemma 5 and taking $u = \log(8T/\varepsilon)$, a union bound over $m \le T$ gives its stated radius simultaneously for every prefix actually reached, with failure probability at most ε. This construction applies conditionally at a random stage start because the compensated processes start at one there and are observed only until the deterministic calendar horizon T. If further visits never occur, their increments remain zero. Thus no recurrence assumption, fictitious future observation, or conditioning on future completion is used.

Probability allocation. For each pair–fast-head index (x, j), apply the stopped form (38) of Lemma 5 with failure probability $\delta/(16SAJ)$. Its conclusion is simultaneous over every reached prefix $m \le T$, so the union over the at most SAJ such sequences costs at most $\delta/16$. A slow head is reset at every stage. Condition on the history immediately before the first owner transition of a particular stage, when its head still equals the reset value U, and apply (38) to the fresh stage recursion with failure probability $\delta/(16SAJT)$ (again simultaneously over all reached prefixes within that stage). There are at most T nonempty stages per pair and hence at most $SAJT$ pair–head–stage instances; this second union costs at most $\delta/16$. The stage-start conditioning is legitimate because this first owner visit is selected before its transition, all stage targets are frozen before their fresh beta draws, and no update of this stage has yet occurred.

To spell out the random-block union, enumerate first owner visits of stages chronologically and pad with cemetery starts, exactly as in the proof of Lemma 6. Write ζ_r for such a start, set $I_r = \mathbf{1}\{\zeta_r \le T\}$, and take $\mathcal{R}_r = \mathcal{F}_{\zeta_r}$ on actual starts, with the terminal sigma-field for cemetery starts. The owner, stage parameters, and I_r are known in this record. In particular, I_r does not ask whether an already opened stage will receive a future visit. The conditional failure probability for any fixed head is at most $\delta/(16SAJT)$ on $\{I_r = 1\}$. Therefore

$$\mathbb{P}\left(\bigcup_{r,j} \{I_r = 1, \mathcal{C}_{r,j}^c\} \right) \le \sum_{r,j} \mathbb{E}\big[I_r \mathbb{P}(\mathcal{C}_{r,j}^c \mid \mathcal{R}_r, I_r = 1) \big] \le \frac{\delta}{16}.$$

The same tower argument applies to the slow transition-noise blocks below. Thus neither union conditions on the random future endpoint of a stage, and the beta deviations of every observed prefix of an unfinished last stage are covered. Slow transition-noise bounds are needed only for completed stages.

For the transition terms, use (37), the stopped form of Lemma 12, for every fast prefix (x, m) with failure probability $\delta/(16SAT)$. For a slow stage (x, q) use $m = M_q$ and the same per-block failure probability, conditional on its starting history; stages with $M_q > T$ cannot finish and require no such bound. Fast weights are deterministic for fixed (x, m), and slow weights are deterministic once the stage parameters are fixed. In each case the failure event includes existence of the required prefix,

rather than conditioning on its existence. There are at most SAT fast prefixes and at most SAT nonempty slow blocks, so these two unions cost at most $\delta/16$ each. Thus all four concentration families fail with probability at most $\delta/4$, which is stronger than the claimed $\delta/2$ envelope probability. Quantitatively, the largest logarithm in the beta bounds is

$$\log \frac{128SAJT^2}{\delta} \le 2\ell + \log J,$$

and weighted Bernstein uses at most $\log(32SAT/\delta) \le \ell$. Since $J = \lceil c_J \ell \rceil$, $\ell \ge \log 256$, and c_J is universal, $\log J \le C\ell$. Thus every concentration logarithm is at most a universal multiple of ℓ. Enlarging the universal constant C absorbs the resulting numerical factors. On this joint event, the preceding fast decomposition holds for every $j = 1, \dots, J$ with the same right-hand side: its bootstrap and transition terms use the deterministic mean weights, while the head-specific rate deviation is bounded uniformly by b_m. Taking the maximum over j therefore gives (35). For a fixed completed slow stage and a fixed temporary head, the same event bounds that head by the displayed slow right-hand side. The union over all stage–head instances was already paid for above, so the bound holds simultaneously for every head in the stage. Since the commit is their maximum, taking the maximum now gives (36). Thus neither maximum introduces an additional random term or an unaccounted failure probability. The fast initialization is at most $2H_\gamma \bar{W}_m^0$. For the slow update, the exact prior mass is $n_0^\flat/(n_0^\flat + M)$ and each data mean weight is $1/(n_0^\flat + M)$, producing (34). The nonnegativity assertion for the slow error is supplied separately by Lemma 9, which is why the main proof intersects $\mathcal{E}_{\mathrm{env}}$ with $\mathcal{E}_{\mathrm{opt}}$ rather than charging its failure probability twice. □

H Occurrence-indexed dependency recursion

Remark (Owner convention for occurrence columns). The row indexed by a visit to $x = (s, a)$ contains only occurrences from the slow stage that committed the current entry $\widetilde{Q}^\flat(x)$. Hence every direct coefficient in that row has owner x. A value estimate at a successor state can depend on a different action pair only through the error variable d_j attached to the later occurrence time j; this dependence is represented by the recursive term $\gamma A d$, not by inserting the older pair's sample column directly into the current row. This convention permits a pairwise column bound that remains valid after summing over owners.

Let

$$D_t = Q_t(s_t, a_t) - Q^*(s_t, a_t), \qquad d_t = D_t + \bar{\eta} H_\gamma.$$

For a pair $x = (s, a)$, let $N_T(x) = \sum_{t=1}^T \mathbf{1}\{(s_t, a_t) = x\}$ be its number of visits through time T; hence $\sum_x N_T(x) = T$. For matrices, $\|\boldsymbol{A}\|_1$ denotes the induced ℓ_1 operator norm, equivalently $\|\boldsymbol{A}\|_1 = \max_j \sum_t |\boldsymbol{A}_{tj}|$. For vector notation below, define $E_j^\flat = \widetilde{V}_j^\flat(s_j) - V^*(s_j)$ and write $\boldsymbol{E}^\flat = (E_1^\flat, \dots, E_T^\flat)^\top$.

Lemma 14 (Selected-action and slow-value relations)**.** *On the optimism event,*

$$0 \le g_t \le d_t, \qquad \widetilde{V}_t^\flat(s_t) - V^*(s_t) \le d_t + 2\bar{\eta} H_\gamma. \tag{39}$$

Proof. Let $a_t^* \in \arg\max_a Q^*(s_t, a)$. Greediness and (31) give

$$Q_t(s_t, a_t) \ge Q_t(s_t, a_t^*) \ge (1 - \bar{\eta}) V^*(s_t).$$

Therefore $D_t \ge g_t - \bar{\eta} V^*(s_t) \ge g_t - \bar{\eta} H_\gamma$, which proves the first claim.

Let $b_t \in \arg\max_a \widetilde{Q}_t^\flat(s_t, a)$. Greediness of the policy action and nonnegativity of the fast head give

$$Q_t(s_t, a_t) \ge Q_t(s_t, b_t) \ge (1 - \eta_t(s_t, b_t)) \widetilde{V}_t^\flat(s_t).$$

Writing $\widetilde{V}_t^\flat(s_t) = V^*(s_t) + E_t^\flat$ and $Q_t(s_t, a_t) = V^*(s_t) - g_t + D_t$ yields

$$(1 - \eta_t(s_t, b_t)) E_t^\flat \le D_t - g_t + \eta_t(s_t, b_t) V^*(s_t) \le D_t + \bar{\eta} H_\gamma.$$

Because $0 \le E_t^\flat \le 2H_\gamma$, $E_t^\flat \le D_t + 3\bar{\eta} H_\gamma = d_t + 2\bar{\eta} H_\gamma$. □

For a slow target observed at time i, call it *exceptional* when the transition is a self-loop and the update at i completes a slow stage. There is at most one exceptional sample per completed stage, and we charge its downstream bootstrap contribution directly. For every other slow sample put

$$\sigma(i) = i + 1. \tag{40}$$

The frozen target reads $\widetilde{V}_i^\flat(s_{i+1})$. If the transition is not a self-loop, the update at i changes values only at s_i; if it is a self-loop but the sample is not exceptional, no slow commit occurs. Consequently, in both cases,

$$\widetilde{V}_i^\flat(s_{i+1}) = \widetilde{V}_{\sigma(i)}^\flat(s_{\sigma(i)}).$$

If this sample influences the policy value used at a later visit t to the same pair, then $\sigma(i) < t$. In the non-self-loop case the trajectory must leave the state before returning. In the self-loop case a nonexceptional sample cannot enter a committed slow value until a later stage-ending visit.

Call every visit in the initial slow stage, and the first visit of every later slow stage, an *exceptional visit*. The initial stage has no preceding commit. The synchronized commit is available to the next action selection. We retain the initial-stage visits and, conservatively, stage-boundary visits as exceptional charges; this only enlarges the bound. Their total number is at most

$$CSAL \log(2T). \tag{41}$$

Indeed, for a fixed pair x, the initial stage contributes at most $M_0 = L$ visits, while every later started stage contributes at most one first-visit boundary. Lemma 15 bounds the number of later stages by $CL \log(2T)$; summing over the SA pairs gives (41) after enlarging C.

For the counting statements below, a stage is called *started* (or *nonempty*) only after it has received its first visit. Thus resetting a counter after a commit opens and initializes the next stage but does not count that stage as started until the pair is visited again. In particular, a pair with no visits has no started stage.

Lemma 15 (Stage arithmetic)**.** *For every pair with N visits,*

$$\sum_{q:M_q \text{ started}} \sqrt{M_q} \le CL\sqrt{N}, \qquad \frac{M_{q+1}}{M_q} \le 1 + \frac{2}{L}, \qquad \#\{q : M_q \text{ started}\} \le CL \log(2T). \tag{42}$$

Proof. If $N = 0$, no stage is started and all three claims are immediate. Assume therefore that $N \ge 1$. Put $x_q = L(1 + 1/L)^q$. Since $L \ge 1$ and $x_q \ge L$, the outer maximum in the definition of M_q is inactive and $M_q = \lfloor x_q \rfloor \ge L$. If $M_q = L$, then $M_{q+1} \le L + 1$. If $M_q \ge L + 1$, integrality and $x_q < M_q + 1$ give

$$M_{q+1} \le M_q + \left\lceil \frac{M_q + 1}{L} \right\rceil \le M_q + \frac{2M_q}{L}.$$

This proves the middle claim. Since $\log(1 + 1/L) \ge 1/(2L)$ for $L \ge 1$, an interval whose endpoints have ratio two contains at most $1 + 2L \log 2 \le CL$ consecutive stage indices. Group stages into the blocks $2^r L \le M_q < 2^{r+1} L$. Each block contains at most CL stages. Indeed, since the block endpoints and M_q are integers, $M_q \le x_q < M_q + 1$ implies $2^r L \le x_q < 2^{r+1} L$ for every index in block r; hence the preceding ratio-two count remains valid despite the floor in M_q. Summing the geometric square roots gives the first claim, and counting the blocks gives the last. More explicitly, within block r, each term is at most $\sqrt{2^{r+1} L}$ and there are at most CL terms, so its contribution is at most $CL\sqrt{2^{r+1} L}$. Let $q_\star$ be the largest started-stage index and let $r_\star$ be the block containing $q_\star$. If $r_\star \ge 1$, the preceding stage $q_\star - 1$ has already been completed, so $N \ge M_{q_\star - 1}$ and the ratio bound gives $M_{q_\star} \le (1 + 2/L) M_{q_\star - 1} \le 3N$. Hence

$$\sum_{r=0}^{r_\star} CL\sqrt{2^{r+1} L} \le C' L \sqrt{2^{r_\star} L} \le C'' L\sqrt{N}.$$

If $r_\star = 0$, then either only the initial stage has started, in which case $\sum_q \sqrt{M_q} = \sqrt{L} \le L\sqrt{N}$, or a later stage has started, in which case $N \ge L$ and there are at most CL started stages in this block; therefore $\sum_q \sqrt{M_q} \le CL\sqrt{2L} \le C' L\sqrt{N}$. Finally, if only the initial stage is started there is one nonempty block. Otherwise $q_\star \ge 1$, the preceding stage is completed, and the same ratio argument gives $M_{q_\star} \le 3N$, so $2^{r_\star} L \le M_{q_\star} \le 3N$ and the number of nonempty blocks is at most

$$1 + r_\star \le 1 + \log_2(3N/L) \le C \log(2T),$$

where the last inequality uses $1 \le N \le T$ and absorbs the cases $3N/L < 1$ into the preceding one-block alternative. Multiplying by the CL stage indices per block proves the last assertion, including an unfinished final stage. □

Lemma 16 (Slow occurrence matrix)**.** *Keeping the full T-coordinate interaction-time indexing $1, \dots, T$, set the rows corresponding to exceptional visits to zero and charge exceptional slow samples directly. Then there is a strictly lower-triangular, entrywise nonnegative $T \times T$ matrix A such that*

$$\|\boldsymbol{A}\|_1 \le 1 + \frac{2}{L}, \qquad \sum_j \boldsymbol{A}_{tj} \le 1. \tag{43}$$

No coordinate is deleted: a column j whose visit is exceptional is retained, since later dependencies may enter through it, while row j of A is zero and the corresponding error d_j is handled by the nonrecursive charge r_j of Lemma 17. For every nonexceptional visit, the remaining regular-sample slow bootstrap part of its policy-Q error, after the exceptional-sample contribution has been assigned to the nonrecursive charge, is at most

$$\gamma \sum_{j<t} \boldsymbol{A}_{tj} \{\widetilde{V}_j^\flat(s_j) - V^*(s_j)\}. \tag{44}$$

Proof. Suppose the slow entry used at a nonexceptional visit to x was committed from stage q. For every $i \in \mathcal{I}_q(x)$, put $1/(n_0^\flat + M_q)$ in column $\sigma(i)$ of that row, multiplied by the outer mixture coefficient $1 - \eta \le 1$, except that an exceptional sample is omitted and charged directly. If $\sigma(i) = T + 1$, the sample cannot be read by any update within the analyzed horizon and is omitted as well. The discussion after (40) makes the matrix strictly lower triangular, and the row sum is at most $M_q/(n_0^\flat + M_q) \le 1$.

To verify strictness, let c be the commit time and let $t > c$ be a nonexceptional row reading that commit. If $i < c$, then $\sigma(i) = i + 1 \le c < t$. If $i = c$ and $s_{c+1} \ne s_c$, then $\sigma(i) = c + 1$; the first post-commit owner visit is exceptional, so every nonexceptional row using the commit has $t > c + 1 = \sigma(i)$. If $i = c$ and $s_{c+1} = s_c$, the sample is exceptional, omitted from A, and charged directly; only in this case can the target value change at its own successor occurrence. Finally, if $i = T$, then $\sigma(i) = T + 1$ lies beyond the analyzed vector and no row $t \le T$ can read a commit containing that sample. These cases prove $A_{tj} > 0 \Rightarrow j < t$, including stage boundaries, self-loops, and the horizon endpoint.

A regular sample from stage q is present only in that commit. After discarding exceptional visits, that commit enters at most M_{q+1} policy-Q updates before it is replaced. Hence the total mass in the sample's column is at most $M_{q+1}/(n_0^\flat + M_q) \le M_{q+1}/M_q \le 1 + 2/L$. The first owner visit in stage $q + 1$ is exceptional and contributes no row; the coarser bound M_{q+1} is intentional and remains valid under horizon truncation. The map $i \mapsto \sigma(i) = i + 1$ is injective, so merging sample columns into occurrence columns does not increase the column sum. Equation (44) now follows from (36). Each row at time t is zero if t is exceptional. Otherwise, it contains the sample coefficients of the committed stage read at time t. A sample contributes at most once per row and only while its committed entry remains current. A column with $\sigma(i) = T + 1$ cannot appear in a row indexed by $t \le T$. Attach to each occurrence column $j = \sigma(i)$ its owner $x_i = (s_i, a_i)$ and the stage q_i whose commit contains sample i. The associated value is $E_j^\flat = \widetilde{V}_j^\flat(s_j) - V^*(s_j)$ with $s_j = s_{i+1}$: "owner" identifies which committed Q-entry reuses the coefficient, whereas the column value is the successor-state error appearing in sample i's Bellman target. Equation (40) links the owner sample to its successor-state value. Column j occurs only in rows corresponding to later visits of the same owner x_i, while that stage-q_i commit remains current; there are at most M_{q_i+1} such visits. Indeed, each row has one visited owner, a commit changes only that owner's entry, and a successor-state value is read only on a later transition. A later visit of a different predecessor pair does not reuse column j: it creates its own target occurrence and hence a new column, whose dependence on j is represented in a later row of the strictly lower-triangular recursion. Cross-state predecessor reads are therefore represented by new occurrence columns. The row bound of one and column bound of $1 + 2/L$ remain valid across all predecessor pairs. Column j is retained when the visit at time j is exceptional. Only row j of A is set to zero; the variable d_j remains and is bounded by the nonrecursive charge $r_j = 3H_\gamma$ in Lemma 17. This preserves dependencies entering through exceptional occurrence columns.

If an exceptional sample is omitted, its coefficient in the stage-q commit is $1/(n_0^\flat + M_q)$. The commit can enter at most M_{q+1} subsequent nonexceptional policy updates, each with outer coefficient at most one. Because its target excess is at most $2H_\gamma$, its entire downstream contribution is therefore at most

$$\frac{2H_\gamma M_{q+1}}{n_0^\flat + M_q} \le 2H_\gamma \left(1 + \frac{2}{L}\right).$$

Thus charging one such sample per completed stage costs $O(SAH_\gamma L \log(2T))$ in total. This charge is only for the exceptional sample's direct coefficient in its owner's committed entry. If the resulting committed state value is later read by a predecessor pair, that read belongs to the predecessor's own target occurrence and remains in the recursive matrix term. □

Lemma 17 (Nonrecursive charge)**.** *On the intersection of the preceding high-probability events, there are nonnegative r_t such that*

$$d_t \le r_t + 3\bar{\eta}H_\gamma + \gamma(\boldsymbol{A}\boldsymbol{d})_t \tag{45}$$

and

$$\sum_{t=1}^{T} r_t \le CH_\gamma L\{\ell^2\sqrt{SAT} + SA\ell^3\}. \tag{46}$$

Proof. At an exceptional visit set $r_t = 3H_\gamma$; this dominates d_t because $Q_t \le U$. Consider a nonexceptional visit. Insert (35)–(36) into the exact policy-mixture identity

$$D_t = \eta_t\left\{\max_j \widehat{Q}_t^j(s_t,a_t) - Q^*(s_t,a_t)\right\} + (1-\eta_t)\left\{\widetilde{Q}_t^\flat(s_t,a_t) - Q^*(s_t,a_t)\right\}.$$

The mixture identity holds pathwise. In the fast envelope, the bootstrap coefficients have total mass at most one and $\widehat{V}_i(s) - V^*(s) \le U = 2H_\gamma$, so its entire bootstrap term is at most $2H_\gamma$ before multiplication by η_t. The slow bootstrap terms are the only terms retained recursively and are placed in the row of A specified by Lemma 16. The resulting $2H_\gamma\eta_t$ fast charge is assigned to r_t. The fast initialization, both confidence radii, the slow prior mass, and the downstream contribution of every exceptional slow sample are also assigned to r_t. To define the charge for each row, let m_t be the stored fast counter $N(s_t,a_t)$ immediately before action selection at time t, equivalently the number of visits to the selected pair completed strictly before time t. Accordingly, the stored fast entry read at time t is the iterate after exactly m_t updates; the transition at time t produces update m_t+1 only after the action has been selected. On a nonexceptional row a committed slow stage necessarily exists; write $M_t = M_q$ for its length, let c_t denote the commit time of that completed slow stage, that is, of the commit that produced the slow entry currently read for the selected pair (s_t,a_t), and let

$$e_t := \mathbf{1}\{s_{c_t+1} = s_{c_t}\}.$$

Thus e_t records whether the creation commit of the currently read slow entry ended with a self-loop; it is not an indicator of whether the transition at time t is a self-loop, nor is c_t a single commit time for the whole table. Rows before the first commit are among the exceptional rows charged by definition. Here and below η_t has precisely the selected-pair meaning in (4); in particular it is the coefficient actually stored in $Q_t(s_t,a_t)$, rather than the rate of an unfinished stage inferred retrospectively. Set, on a nonexceptional row,

$$\begin{aligned} r_t = {} & 2H_\gamma\eta_t + 2H_\gamma\eta_t\bar{W}^0_{m_t} + \eta_t b_{m_t} \\ & + (1-\eta_t)(b^\flat_{M_t} + p^\flat_{M_t}) + (1-\eta_t)e_t\frac{2\gamma H_\gamma}{n_0^\flat + M_t}. \end{aligned} \tag{47}$$

If the current slow entry was committed from stage q, it is read while stage $q+1$ is being collected. The first owner visit after that commit is exceptional; after this visit the persistent entry is recomputed with rate $(M_{q+1}+1)^{-1/2}$. Hence every subsequent nonexceptional row reading the stage-q commit has exactly this stored rate. This justifies charging the single boundary mismatch separately and using the stagewise rate in the summation below. The first term bounds the complete fast bootstrap sum, since its mean weights have mass at most one and every fast value excess is at most $2H_\gamma$; retaining the factor γ would only make it smaller. The last term is the omitted exceptional target's direct coefficient in this particular row. Summing that last term over all rows that read the same

commit gives exactly the aggregate estimate in Lemma 16. A later predecessor read of the resulting state-value error is represented by a new occurrence column in A; it is not another direct copy of this sample. For an exceptional row retain $r_t = 3H_\gamma$. Hence all omitted terms are charged without double counting. The only recursive part left is (44); row by row, the preceding substitution therefore gives the explicit intermediate inequality

$$D_t \le r_t + \gamma(\boldsymbol{A}\boldsymbol{E}^\flat)_t. \tag{48}$$

For an exceptional row this same display holds because that row of A is zero and $D_t \le U - Q^*(s_t, a_t) \le 2H_\gamma < r_t$. Since the row-sum bound in Lemma 16 gives $\boldsymbol{A}\mathbf{1} \le \mathbf{1}$ coordinatewise, the second inequality in the display below uses that row-sum property. Lemma 14 gives

$$\gamma\boldsymbol{A}\boldsymbol{E}^\flat \le \gamma\boldsymbol{A}\boldsymbol{d} + 2\gamma\bar{\eta}H_\gamma\boldsymbol{A}\mathbf{1} \le \gamma\boldsymbol{A}\boldsymbol{d} + 2\bar{\eta}H_\gamma\mathbf{1}.$$

Adding the $\bar{\eta}H_\gamma$ in $d_t = D_t + \bar{\eta}H_\gamma$ proves (45).

It remains to verify (46). Lemmas 3 and 15, integral comparison, and Cauchy–Schwarz give the following bounds. Let $\mathcal{Q}_x^{\mathrm{comp}}$ denote the indices of completed slow stages for pair x.

$$\sum_{t \text{ nonexceptional}} \eta_t \le CL\sqrt{SAT},$$

$$\sum_x \sum_{m=0}^{N_T(x)-1} b_m \le CH_\gamma\{\sqrt{LSAT\ell} + SAL\ell^2\},$$

$$\sum_x \sum_{q \in \mathcal{Q}_x^{\mathrm{comp}}} M_{q+1} b^\flat_{M_q} \le CH_\gamma L\{\ell\sqrt{SAT} + SA\ell^3\},$$

$$\sum_x \sum_{q \in \mathcal{Q}_x^{\mathrm{comp}}} M_{q+1} p^\flat_{M_q} \le CSAH_\gamma L\ell^3.$$

The four displayed lines correspond, in order, to the terms proportional to η_t (including the fast bootstrap), the fast radius, the slow radius, and the slow prior in the explicit definition (47). The fast initialization, exceptional visits, and exceptional-sample coefficient are bounded separately below. The factors M_{q+1} in the last two lines bound the number of rows that read a completed stage-q entry. If the next stage $q+1$ is unfinished at time T, its visits still read the preceding stage-q commit and are included among these at most M_{q+1} rows; the unfinished stage's own samples create no additional slow entry. Its fast updates are covered by the first two lines. This also covers $N_T(x) = 0$, for which all corresponding sums are empty. For the first line, Lemma 15 gives, pairwise, $\sum_q M_q/\sqrt{M_q+1} \le CL\sqrt{N_T(x)}$; this applies to the nonexceptional visits, whose stored mixing rate is the rate of their current stage. The first visit after a stage commit is already an exceptional charge, and its possibly stale preceding rate is absorbed there. Summing over pairs and using Cauchy–Schwarz therefore yields $\sum_{t \text{ nonexceptional}} \eta_t \le CL\sqrt{SAT}$. For the fast-radius sum, integral comparison (with an empty sum when $N = 0$) gives, for $N \ge 1$,

$$\sum_{m=0}^{N-1} \sqrt{\frac{L\ell}{L+m}} \le C\sqrt{L\ell N}, \qquad \sum_{m=0}^{N-1} \frac{L\ell}{L+m} \le CL\ell\log(2T).$$

Summing the first inequality over pairs and using $\sum_x \sqrt{N_T(x)} \le \sqrt{SAT}$, while using $\log(2T) \le C\ell$, gives the second line. For a slow stage, $M_{q+1} \le CM_q$ and Lemma 15 imply the following bound; to make its two terms explicit, use the definition of $b^\flat_M$, $M_{q+1} \le CM_q$, and $n_0^\flat + M_q \ge M_q$:

$$M_{q+1} b^\flat_{M_q} \le CH_\gamma\{\sqrt{\kappa^\flat \ell M_q} + \kappa^\flat\ell\}.$$

Lemma 15 therefore gives

$$\sum_q M_{q+1} b^\flat_{M_q} \le CH_\gamma\{L\sqrt{\kappa^\flat\ell}\sqrt{N_T(x)} + L\kappa^\flat\ell\log(2T)\}.$$

Since $\kappa^\flat = c_\flat\ell$ and $\log(2T) \le C\ell$, this is the displayed pairwise bound. Summing it and using $\sum_x \sqrt{N_T(x)} \le \sqrt{SAT}$ proves the third line. For the prior term, the exact definition of $p^\flat_M$ and $M_{q+1} \le CM_q$ give

$$M_{q+1} p^\flat_{M_q} = \frac{2H_\gamma n_0^\flat M_{q+1}}{n_0^\flat + M_q} \le CH_\gamma n_0^\flat.$$

There are at most $CL\log(2T)$ completed stages for each pair, and $n_0^\flat = O(\ell^2)$, hence the sum over all pairs is $O(SAH_\gamma L\ell^3)$, proving the fourth line. The factors M_{q+1} bound the number of owner visits at which each committed stage remains current. The fast prior satisfies, with $a = L+1$ and $b = n_0 = 1$,

$$\sum_{m=0}^{\infty} \bar{W}_m^0 = \sum_{m=0}^{\infty} \frac{(1)_m}{(L+2)_m} = {}_2F_1(1,1;L+2;1) = \frac{L+1}{L} = 1 + \frac{1}{L},$$

by Gauss' identity. Thus its total contribution is $O(SAH_\gamma)$. The exceptional-visit and exceptional-sample charges are $O(SAH_\gamma L\log(2T))$ by (41). All these terms are absorbed by the right-hand side of (46). Explicitly, $\ell \geq \log 256 > 1$, $L \geq 1$, and $SA \geq 1$ imply

$$H_\gamma L\sqrt{SAT} \leq H_\gamma L\ell^2\sqrt{SAT}, \qquad H_\gamma\sqrt{LSAT\ell} \leq H_\gamma L\ell^2\sqrt{SAT},$$
$$SAH_\gamma L\ell^2 \leq SAH_\gamma L\ell^3, \qquad SAH_\gamma L\log(2T) \leq C\,SAH_\gamma L\ell^3.$$

The remaining $O(SAH_\gamma)$ initialization term is dominated by the last right-hand side as well. The resulting constant C is independent of S, A, T, γ, and δ. □

I Experimental specification, statistics, and availability

This section records the settings actually used by the submitted raw runs. They are practical implementations, not executions of the sufficient constants in Theorem 1.

I.1 Criterion-aware comparison

Table 3 separates each paper's native criterion from its leading scale under the reward-based Liu–Su criterion. Polylogarithms and lower-order terms are suppressed. The UCBVI-γ entry uses the comparison in Appendix A.2 of He et al. (2020); it is not obtained by equating the two native regret definitions. The last column always denotes the reward-based criterion in equation (6).

Result	Native leading term	Liu–Su leading scale
Liu–Su Double Q (Liu & Su, 2020)	$\widetilde{O}(H_\gamma^{3/2}\sqrt{SAT})$	same (original upper)
UCBVI-γ (He et al., 2020)	$\widetilde{O}(H_\gamma^{3/2}\sqrt{SAT})$ (value regret)	$\widetilde{O}(H_\gamma^{1/2}\sqrt{SAT})$
Zamir–Zurek–Chen (Zamir et al., 2026)	$\widetilde{O}(\sqrt{SA\,\mathrm{Var}})$ (γ-regret)	$\widetilde{O}(H_\gamma^{1/2}\sqrt{SAT})$ (worst case)

Table 3: Leading $\sqrt{SAT}$ terms at a common reward-based metric. The UCBVI-γ translation multiplies its native value-regret scale by $1-\gamma$ and retains endpoint, martingale, regime, and lower-order terms. Zamir–Zurek–Chen directly treat Liu–Su-style γ-regret; their optimal variance-dependent guarantee retains the minimax worst-case scale while adapting to variance/span and benign instances.

I.2 Exact environments and randomization

Maze-25. Rows and columns are $0,\ldots,24$, with state index $s = 25r + c$; the canonical actions $0,1,2,3$ are up, down, left, and right. The frozen map has start $(0,0)$, goal $(24,24)$, and 568 internal blocked edges. It was generated by `randomized-dfs-plus-openings-v1` with map seed 20005 and eight extra openings; the generation constraints were shortest-path length 90–120, minimum dead-end length 5, at least three dead ends, and alternative-path ratio at least 1.40. For a non-goal state, the intended direction has probability 0.9 and its two perpendicular directions probability 0.05 each. An out-of-bounds or blocked move self-loops. Raw rewards are -0.1 for a self-loop, 1 for entering the goal, and 0 otherwise; the stored reward is $(r+0.1)/1.1$, hence $0, 1/11, 1$. At the goal, the next transition is deterministically to the start with raw reward 0 and consumes no transition RNG draw. There is one initial reset and no episodic reset. Base seeds are $0{:}29$ under master seed 2026082801. Each seed has one PCG64DXSM-derived permutation of the four external action labels, shared by all methods; environment draws are visit-indexed by state, canonical action, and channel. The run length is $2{,}000{,}000$ and there are 200 checkpoints.

Chain-50. States are $0, \dots, 49$ and the initial state is 0. Semantic actions attempt left or right; the requested direction occurs with probability 0.9 and the opposite direction with probability 0.1. Motion is clipped at the two boundaries, producing a self-loop. The deterministic reward is attached to the pre-transition state and is action independent:

$$r(s) = 0.05 + 0.15(1 - x) + 0.95x^4, \qquad x = s/49.$$

Thus $r(0) = 0.20$, $r(49) = 1$, and all intermediate rewards are positive. The chain never terminates or resets after initialization. Each of base seeds 0:29 is run with both external-to-semantic permutations $(0, 1)$ and $(1, 0)$ under master seed 2026083101. The environment and algorithm seeds are derived from the base seed but not from the permutation; the two permutations are averaged within base seed before inference. The run length is 500,000 and there are 250 checkpoints.

I.3 Practical algorithms and tuning

D-RandQ and Slow-only. Both tasks use $U = 100$, $J = 16$, fast $(\kappa, n_0) = (1, 1)$, slow $\kappa^\flat = \ell$, $n_0^\flat = \lceil 2\ell^2 \rceil$, stage lengths $M_q = \lfloor L(1 + 1/L)^q \rfloor$, and completed-stage mixing $\eta_q = (M_q + 1)^{-1/2}$. For Maze-25, $(L, \ell, n_0^\flat) = (2361, 30.18046628685412, 1822)$; for Chain-50 these are $(2223, 25.57529610086603, 1309)$. Slow-only makes three joint changes relative to D-RandQ: it removes the fast heads and the fast–slow mixture, so the executing policy Q follows the committed slow table $\widetilde{Q}^\flat$ instead of a per-step mixture with the fast maximum; it therefore refreshes that policy only when a stage is committed rather than at every visit to the pair; and it replaces the fast bootstrap source $\widehat{V}$ by the committed slow value $\widetilde{V}^\flat$. Target freezing is *not* a difference: exactly as in Algorithm 1, both variants freeze their targets before any beta draw. The commit rule is likewise unchanged: both variants commit the maximum of the temporary heads only at a stage boundary and then reset those heads to U. D-RandQ-mean replaces every beta draw by its conditional mean; identically initialized heads consequently remain identical. No ablation variant was separately tuned.

QSSA. The local implementation follows Algorithms 1–2 of Ji & Li (2023) (arXiv:2305.15546v2), with observed rewards and the continuing trajectory. It sets $H = \lceil 2/(1 - \gamma) \rceil = 200$, $\iota = \log(SAT/\delta)$, $\eta_n = (H + 1)/(H + n)$; initializes $Q, Q^{\mathrm{UCB}}, Q^{\mathrm{ref}}, Q^{\mathrm{lazy}}, Q^{\mathrm{marker}}, V, V^{\mathrm{ref}}$ to H_γ, lower tables and moment accumulators to zero; and applies the paper's UCB/LCB, reference-advantage, and variance updates. The elementary bonus is $c_b\sqrt{\iota/((1 - \gamma)^3 n)}$; the same c_b multiplies the variance, reference, and $n^{-3/4}$ bonus components in our implementation. The current pair is buffered, a pending lazy-table switch is applied on the next observation, and a switch is triggered when accumulated decrease exceeds H_γ. The reference value remains open while $V - V^{\mathrm{LCB}} > 3$ and is updated once more when that condition first fails. There is no model-based planning or numerical tolerance. Here c_b is treated as a practical multiplier rather than a theoretically specified constant.

UCBVI-γ. The local empirical-reward implementation follows Algorithm 1 and equations (4.1)–(4.3) of He et al. (2020) (arXiv:2010.00587v3). It initializes Q, V to H_γ and counts and empirical reward/transition statistics to zero. With

$$U_{\log} = \log\frac{40SAT^3 \log^2 T}{\delta(1 - \gamma)^2}, \qquad \beta = S^2A^2U_{\log}^5,$$

the code uses the Bernstein bonus

$$\sqrt{\frac{8U_{\log} \operatorname{Var}_{\widehat{P}}(V)}{n}} + \frac{8U_{\log}}{(1 - \gamma)n} + \sqrt{\frac{8}{n}\widehat{P} \min\left\{\frac{100\beta}{(1 - \gamma)^5(N_s \vee 1)}, H_\gamma^2\right\}}.$$

The practical multiplier c_{VI} multiplies γ times this whole bonus.

When the empirical model and the value table are updated. The empirical transition counts, state counts, and empirical reward mean are updated after *every* interaction. Planning is lazy. A planning event is triggered only when the updated pair count becomes a power of two $(1, 2, 4, 8, \dots)$, so a pair with n visits triggers at most $\lceil \log_2 n \rceil$ events and the total number of events is at most $O(SA \log T)$. Each planning event performs exactly one global Bellman sweep over all pairs with at least one sample, and the sweep is computed throughout from the value table as it stood *before* the sweep. The per-pair update is the one-step update of equation (4.2) of He et al. (2020), evaluated from that pre-sweep table, and the code stores the smaller of the candidate and the entry's current value, so the stored table is non-increasing over time.

Limitation of the baseline implementation. Our practical UCBVI-γ baseline therefore triggers planning at count-doubling events rather than recomputing the value update after every interaction. This choice reduces the computational cost of the model-based baseline and should be understood as a practical implementation variant rather than an exact reproduction of the per-interaction update schedule analyzed by He et al. (2020). We refer to it as our practical UCBVI-γ implementation throughout. The plotted action gap is evaluated separately, by value iteration to residual 10^{-12} and at most 10^7 sweeps on the true model; this offline evaluation is independent of the baseline's internal planning.

I.4 Certified versus practical ensemble size

The guarantee of Theorem 1 holds only under the certified parameter choices of Appendix A.2.1, which differ from the settings actually run above. The proof uses

$$J = \lceil c_J \ell \rceil, \qquad c_J = \frac{2}{p_0}, \qquad p_0 = \frac{1}{2}\overline{\Phi}(1),$$

together with a worst-case per-head anti-concentration lower bound and a union bound over all possible slow commits; it also requires $\kappa^\flat = c_\flat \ell$ with $c_\flat = 8$. Consequently the certified ensemble size is intentionally conservative and can be much larger than the practical choice $J = 16$ used here: evaluating the certified formula at the ℓ values recorded above gives $J = 761$ for Maze-25 and $J = 645$ for Chain-50, while the practical runs set $\kappa^\flat = \ell$, corresponding to $c_\flat = 1$ rather than the certified $c_\flat = 8$. The experiments therefore use $J = 16$ as a practical hyperparameter and are not covered by the simultaneous optimism guarantee of Lemma 9; they do not verify the theorem under its certified parameter regime.

On non-worst-case blocks the actual per-head optimism probability may be larger than the uniform lower bound p_0, because the proof replaces the realized empirical-likelihood quantity by its worst-case confidence-bound value; this suggests that the worst-case analysis is conservative here. However, the present analysis does not establish an instance-dependent guarantee showing that $J = 16$ suffices.

Two further remarks separate this theory–practice gap in constant choices from any contradiction of the theorem. First, the certified constants are deliberately worst-case and the present proof may be quantitatively loose; the reported runs are not intended to instantiate the certified algorithm numerically, nor to verify the finite-sample upper bound. Second, the optimistic-initialization ablation in Figure 3(c) compares the practical $U = 100 = H_\gamma$ with the certified $U = 2H_\gamma = 200$: the certified value does not perform better on Maze-25, and the two curves are essentially coincident at the endpoint. This is a difference in constant choices, not evidence against the theorem.

Sensitivity of the ensemble width was examined only during calibration: the dense-chain sweep of Section I.5 searched $J \in \{2, 4, 6, 8, 10, 12, 16\}$ and selected $J = 12$, which is close to but not equal to the practical $J = 16$ fixed on C-Maze-15. We ran no J sweep on the reported Maze-25 or Chain-50 tasks themselves; a broader J-sensitivity study under the certified regime remains future empirical work.

I.5 Calibration of the practical multipliers

The practical multipliers and the practical ensemble width were selected on a calibration task that is disjoint from the reported benchmarks, using a fixed candidate grid recorded in the experiment configuration and a single selection rule. Table 4 records the grid, the screen, and the outcome. The calibration benchmark `C-Maze-15` is a 15×15 maze generated with map seed 20003 and four extra openings, disjoint from the reported map seeds 20005 (Maze-25) and 21007 (Maze-15). Each candidate was run for 10 calibration seeds over 100,000 interactions; the selection metric was the mean cumulative gap per interaction, `gap_sum`$/T$. Every candidate passed the recorded progress screen (fraction 1.0 of seeds), and the selection rule was the minimum equal-weight mean within-environment rank among progress-eligible candidates, with ties assigned to the smaller value. The frozen outcome is recorded in `outputs/calibration/frozen_parameters.json`. All three algorithms were searched over three candidate values on the same task under the same rule, so the tuning budget was matched across methods.

Algorithm	Parameter	Candidates	Selected
D-RandQ	J	$16, 32, 64$	16
QSSA	c_b	$0.001, 0.003, 0.01$	0.003
UCBVI-γ	c_{VI}	$0.01, 0.03, 0.1$	0.01

Table 4: Calibration grid on C-Maze-15 and selected values. The selection metric was the mean cumulative gap per interaction over 10 calibration seeds at 100,000 interactions.

C-Maze-15 was also the task used to freeze the parameters of the reported Maze-25 and Chain-50 comparisons, and no other calibration entered those runs. The calibration and the reported runs share the same master seed and, for the first ten entries, the same seed indices, but the environment and algorithm streams are keyed by benchmark and phase, so no reported-run random stream is reused; the reported seed indices 10:29 were not used for selection. The multipliers were fixed before the reported runs, and no reported number was used to select them.

Chain-50 operating point. The reported Chain-50 curves use the values frozen on C-Maze-15—$J = 16$ for D-RandQ and $c_b = 0.003$, $c_{\mathrm{VI}} = 0.01$ for the baselines—with no additional Chain-50-specific retuning, as recorded in `configs/experiment.json` and `configs/experiment_dense_u100.json`. The archive separately contains a dense-chain calibration sweep over `Calibration-Chain-20` and `Calibration-Chain-50` (10 seeds, 50,000 and 100,000 interactions, metric: equal-chain mean of the paired-permutation final cumulative gap per interaction). Its coarse stage searched $J \in \{2, 4, 8, 16\}$ and $c_b, c_{\mathrm{VI}} \in \{0.001, 0.003, 0.01, 0.03\}$; the refined and final stages moved to smaller values and ended at $J = 12$, $c_b = 2.5 \times 10^{-5}$, and $c_{\mathrm{VI}} = 2.5 \times 10^{-5}$, with the selected multipliers at the smallest candidate of the final stage rather than at an interior optimum. Those chain-selected multipliers were not evaluated in the reported full-horizon Chain-50 experiment, so we draw no conclusion about the full-horizon ordering under chain-specific tuning; the reported Chain-50 numbers describe the transferred operating point that was actually run.

Sensitivity. We did not perform an exhaustive hyperparameter study beyond the grids above, and no sensitivity sweep of c_b or c_{VI} exists on the reported Maze-25 or Chain-50 tasks themselves. The bonus multipliers are practical scaling choices rather than theorem-specified constants.

I.6 Paired inference and endpoints

All intervals below are 95% percentile-bootstrap intervals from 10,000 resamples of the 30 base-seed indices. A common resampled index vector is used within each paired comparison. For Chain, each seed-level observation is first averaged across the two action permutations. Intervals are for means, are pointwise for curves, and are not simultaneous confidence bands; no null-hypothesis or equivalence test is reported.

Task	Baseline	Endpoint	D-RandQ − baseline	95% CI
Maze-25	QSSA	reward	37,857.6	[36,116.2, 39,776.1]
Maze-25	UCBVI-γ	reward	27,402.4	[25,673.9, 29,301.6]
Maze-25	QSSA	gap	−38,868.6	[−40,645.3, −37,252.4]
Maze-25	UCBVI-γ	gap	−28,557.6	[−30,362.4, −26,895.9]
Chain-50	QSSA	reward	31,338.5	[31,121.5, 31,560.7]
Chain-50	UCBVI-γ	reward	76,473.0	[74,524.0, 77,686.9]
Chain-50	QSSA	gap	−26,731.2	[−26,916.9, −26,547.3]
Chain-50	UCBVI-γ	gap	−61,048.6	[−61,878.4, −59,702.4]

Table 5: Paired endpoint differences. Positive reward and negative gap differences favor D-RandQ. D-RandQ wins all 30 paired base-seed comparisons in every row. The Chain-50 rows use the practical multipliers transferred from the C-Maze-15 calibration task (Section I.5).

For the mechanism comparisons, D-RandQ minus Slow-only endpoint differences are 24,502.1 [95% CI 22,590.3, 26,600.3] in Maze reward and −24,052.9 [−26,430.7, −21,837.2] in Maze gap; the Chain differences are 230,698.4 [207,231.1, 254,953.0] in reward and −178,886.9 [−195,979.3, −162,173.5] in gap. D-RandQ wins all 30 base seeds in all four comparisons. D-RandQ

minus D-RandQ-mean is unresolved: Maze reward $1{,}096.6$ $[-1{,}210.5, 3{,}297.1]$, Maze gap $-1{,}035.4$ $[-3{,}078.0, 1{,}116.6]$, Chain reward 46.4 $[-47.2, 140.7]$, and Chain gap -35.7 $[-103.1, 31.7]$. Intervals containing zero do not establish equivalence. Moreover, D-RandQ versus Slow-only changes the fast ensemble and mixture, the policy-update schedule, and the bootstrap source ($\widehat{V}$ versus $\widetilde{V}^{\flat}$) together, so it does not identify any one mechanism as causal; the result is architectural. Target freezing is common to both variants. The beta-to-mean intervention also collapses identically initialized heads; it does not isolate sampling variance while holding ensemble diversity fixed. Beta randomization nevertheless remains essential to the present analysis, which relies on Dirichlet anti-concentration over independently drawn weights; what is unresolved above is only its incremental empirical benefit at these budgets. Isolating a single mechanism, for instance by changing only the policy-refresh schedule while holding the ensemble and the bootstrap source fixed, would require variants that we did not run, and we leave that to future work.

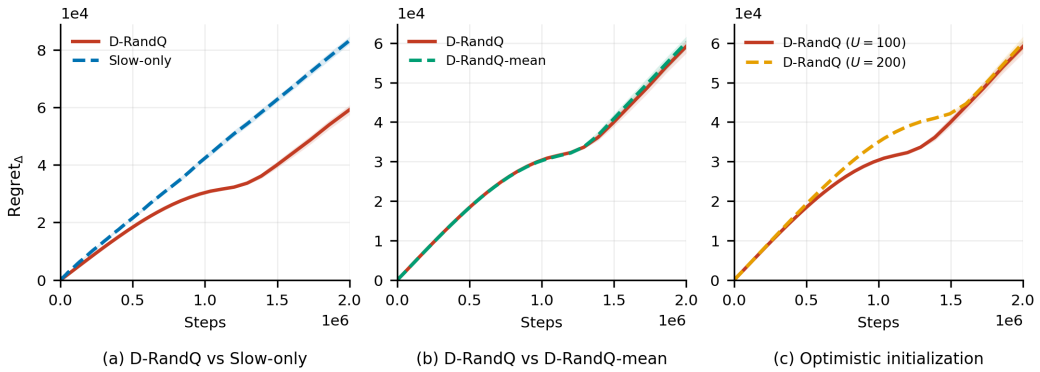


Figure 3: Maze-25 cumulative action-gap ablations over 30 paired base seeds. Shading gives pointwise 95% paired percentile-bootstrap intervals from 10,000 resamples. (a) D-RandQ versus Slow-only; (b) D-RandQ versus D-RandQ-mean; (c) D-RandQ with $U = 100$ versus $U = 200$, with all other settings unchanged.

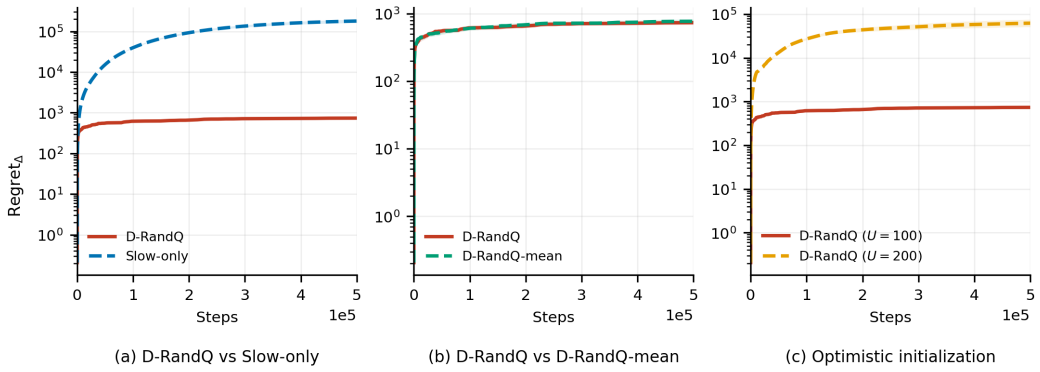


Figure 4: Chain-50 cumulative action-gap ablations on logarithmic vertical scales over 30 paired base seeds, after averaging the two action-label permutations within seed. Shading gives pointwise 95% paired percentile-bootstrap intervals from 10,000 resamples. Panels (a)–(c) use the same comparisons as Figure 3.

J LLM USAGE

Scope of the claim. All formal results stated in this paper—the high-probability action-gap theorem (Theorem 1), its Liu–Su γ-regret corollary (Corollary 1), and every numbered lemma in the main text and in the appendix—were proposed and proved by a large language model (LLM), **GPT-5.6 Sol**,

used in an interactive session. This covers (i) the formulation of the statements themselves, (ii) their derivations, and (iii) the complete proofs, including the randomized-weight algebra, the concentration and optimism arguments, and the occurrence-indexed accounting of the appendix. The authors did not originate these proofs; their role was verification, specified below.

What the model did not produce. The problem formulation, the discounted randomized Q-learning algorithm and its stage schedule, the experimental protocol and its implementation, the reported numbers and figures, and the text of the manuscript are the work of the authors; the contribution of the model is confined to the mathematical content described above. Classical tools invoked inside the proofs, such as Hoeffding's and Bernstein's inequalities, are used as standard results and are not claimed as a contribution of either party. The only other use of the model was language and LaTeX copy-editing of the source, which generated or altered no mathematical content and no reported result.

Verification and responsibility. The central theorem and the key lemmas—in particular those controlling concentration, optimism, and the occurrence-indexed accounting—were checked by hand by the authors, who independently re-derived the main steps. The remaining lemmas, mostly long algebraic or indexing arguments, were checked for consistency with the surrounding results and, where applicable, by finite numerical diagnostics of the stated identities rather than by a complete line-by-line formal audit. No mechanized proof assistant was used. We therefore disclose this use without transferring responsibility for it: the model is not an author of this paper and cannot be held accountable for the results, and the authors take full responsibility for the correctness of every claim and proof, the accuracy of the reported experiments and citations, and the final content of the paper.