Exploring Starts Are Not Enough: Counterexamples and a Fix for Monte Carlo Exploring Starts

Octave Oliviers
Department of Engineering
University of Cambridge
Cambridge, UK
ofao2@cam.ac.uk

Glenn Vinnicombe
Department of Engineering
University of Cambridge
Cambridge, UK
gv103@cam.ac.uk

Abstract

The asymptotic behaviour of Monte Carlo Exploring Starts (MCES) is a long-standing open question in reinforcement learning, even in the tabular setting. We investigate the convergence properties of tabular MCES through examples of persistent suboptimal behaviour. This paper presents new counterexamples for both initial-visit and first-visit MCES and gives a convergence-restoring modification for the initial-visit case. We demonstrate by simulation that stable suboptimal solutions may exist for initial-visit MCES with sample-average updates even when greedy actions are updated more often than non-greedy actions on average. However, scaling learning rates inversely to update frequencies on a state-by-state basis guarantees convergence to optimality. The normalisation requires only conditional action probabilities, avoiding knowledge of the state-sampling distribution. For first-visit MCES with sample-average updates, we show by simulation and by proof that, with positive probability, solutions can also remain trapped away from the optimal action values. This settles a fundamental open problem and shows that exploring starts alone do not guarantee convergence to optimality. More broadly, these results highlight that convergence depends critically on the relative size and frequency of updates applied to different actions, making the choice of learning rates and the balance between exploration and exploitation central to the analysis of MCES and the implementation of scalable Monte Carlo control methods.

1 Introduction

Reinforcement learning aims to train an agent to behave optimally in a certain environment. In this paper, we model this interaction as a finite Markov decision process (MDP) evolving over discrete steps. At time t, the agent observes a state $s_t \in \mathcal{S}$ and selects an action $a_t \in \mathcal{A}(s_t)$ according to a policy $\pi(a_t|s_t)$. The environment then produces a reward r_t and transitions to a new state s_{t+1} according to its dynamics $p(s_{t+1}, r_t|s_t, a_t)$. If π is deterministic in s_t, then $\pi(s_t)$ directly refers to the action a_t where $\pi(a_t|s_t) = 1$.

For a policy π, the action-value function q_π assigns to each pair (s, a) the expected discounted return obtained by taking action a in state s, and then following policy π:

$$q_\pi(s,a) = \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t r_t \;\middle|\; s_0 = s,\ a_0 = a \right].$$

Under standard finite-MDP assumptions, there exists an optimal policy π_* that satisfies $q_{\pi_*}(s,a) = \max_\pi q_\pi(s,a)$ for all (s,a). Moreover, π_* may be chosen to be deterministic and greedy with respect to q_{π_*}, meaning $\pi_*(s) \in \arg\max_{a \in \mathcal{A}(s)} q_{\pi_*}(s,a)$ for all s.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Monte Carlo Exploring Starts (MCES) maintains an action-value estimate q_k and a policy π_k greedy with respect to q_k. At iteration k, the algorithm samples an initial state-action pair (s_0, a_0) according to a distribution σ_k and simulates an episode starting from that pair:

$$\mathcal{E}_k := \Big\{ s_0, a_0, r_0, s_1, a_1, r_1, \ldots : (s_0, a_0) \sim \sigma_k,\ (s_{t+1}, r_t) \sim p(\cdot, \cdot | s_t, a_t),\ a_t = \pi_k(s_t),\ \forall\, t \geq 1 \Big\}.$$

It then updates the estimate q_k at pairs (s, a) appearing in $\mathcal{E}_k$ as

$$q_{k+1}(s, a) = q_k(s, a) + \eta_k(s, a)\big(\tilde{q}_k(s, a) - q_k(s, a)\big), \tag{1}$$

while $q_{k+1}(s, a) = q_k(s, a)$ for non-updated pairs. Here, $\eta_k(s, a)$ is a learning rate satisfying the Robbins–Monro conditions $\sum_k \eta_k(s, a) = \infty$ and $\sum_k \eta_k^2(s, a) < \infty$ for each (s, a), and $\tilde{q}_k(s, a)$ is the observed return from that pair onwards:

$$\tilde{q}_k(s_t, a_t) = \sum_{i \geq 0} \gamma^i r_{t+i} \qquad \forall\, t \geq 0. \tag{2}$$

We consider two update rules. The *initial-visit* (IV) method only updates q_k at the initial pair (s_0, a_0). The *first-visit* (FV) method updates each state-action pair in the episode, using the return from its first occurrence in the episode. We also consider different choices of learning rates. The *sample-average* (SA) rule sets $\eta_k(s, a) = 1/(n_k(s, a) + 1)$, where $n_k(s, a)$ is the number of times (s, a) has been updated before time k. Under this rule, $q_k(s, a)$ is the sample average of the observed returns for (s, a) before time k. Alternatively, the learning rates $\{\eta_k\}_{k \geq 0}$ can be a scalar sequence that is independent of (s, a).

To ensure sufficient exploration, the *exploring starts* condition requires that $\sigma_k(s, a) > 0$ for all k and all (s, a). In order to ensure all state-action pairs are visited infinitely often, we introduce a stricter condition: there exists a constant $\sigma_{\min} > 0$ such that $\sigma_k(s, a) \geq \sigma_{\min}$ for all k and all (s, a).

For both IV and FV, the observed return $\tilde{q}_k$ is an unbiased estimate of q_{π_k} [1]. So, defining the zero-mean perturbation $v_k(s, a) := \tilde{q}_k(s, a) - q_{\pi_k}(s, a)$, equation (1) can be written as

$$q_{k+1}(s, a) = q_k(s, a) + \eta_k(s, a)\big(q_{\pi_k}(s, a) - q_k(s, a) + v_k(s, a)\big). \tag{3}$$

MCES is thus a stochastic approximation to the policy evaluation map associated with the current greedy policy. The dynamics are fixed as long as q_k remains in a region where the same policy is greedy. For a policy π, define this region by

$$Q(\pi) \coloneqq \big\{ q : q(s, \pi(s)) = \max_{a \in \mathcal{A}(s)} q(s, a),\ \forall\, s \in \mathcal{S} \big\}. \tag{4}$$

Discussing the FV/SA variant of MCES, Sutton writes: "It is hard to imagine any RL method simpler or more likely to converge than this ... While this simplest case remains open we are unlikely to make progress on any control method for $\lambda > 0$." [2, p. 13] (Here, λ refers to the parameter in Q(λ), TD(λ) etc. [3].) Sutton and Barto later reinforce the importance of this problem by labelling it as "one of the most fundamental open theoretical questions in reinforcement learning." [4, p. 99]

1.1 Previous work

The Robbins–Monro conditions alone do not prevent IV MCES from converging to suboptimal solutions [5, 6]. In these counterexamples, the sampling distribution is strongly biased towards non-greedy actions (i.e. $\sigma_k(s, a_1) > \sigma_k(s, a_2)$ when $q_k(s, a_1) < q_k(s, a_2)$ and vice versa), as described in Appendix A. In contrast, with uniform sampling distributions over all state-action pairs and scalar learning rates, IV MCES converges almost surely to optimality [7, 8, 9]. For large problems, however, sampling uniformly over the state space is not feasible.

For completeness, we note that MCES also converges to optimality when the MDP exhibits a feedforward structure [6, 10], and that more complex multi-step algorithms manage to recover contraction arguments by combining several multi-step approximations [11], or using a look-ahead mechanism [12]. We focus, however, on the behaviour of tabular MCES for general MDPs as even this simple case remains an open question.

1.2 Contributions

We make three contributions. First, we construct an MDP in which IV/SA MCES is observed to not converge to q_{π_*} even when sampling distributions are biased towards greedy actions. Second, we prove that IV MCES converges to q_{π_*} when scaling learning rates inversely to update frequencies on a state-by-state basis. Unlike earlier convergence results, this normalisation does not require uniform sampling over the state space or knowledge of its sampling probabilities. The proof also does not rely on a contraction argument but uses an extension of the combined stability-ODE method. Third, we show that FV/SA MCES can converge to suboptimal solutions despite strict exploring starts.

2 Counterexample for initial-visit MCES with sample-average updates

Consider the MDP in Figure 1 with $\gamma = 0.9$. Each state has two actions: "move" yields a reward of one while "stay" yields no reward. The policy π_* chooses "move" everywhere, denoted by a_*, while $\bar{\pi}$ chooses "stay", denoted by $\bar{a}$. The following sampling strategy keeps the greedy policy among the six suboptimal policies shown in Figure 1. Each initial-visit update leaves at least one "move" action and one "stay" action unchanged, so neither the optimal policy nor the all-"stay" policy can be reached. The figure illustrates the intended sequence of policy changes. Individual sampled trajectories need not follow this sequence indefinitely and can instead remain permanently at a suboptimal policy.

1. When $q_k \in Q(\pi_1)$, start each episode in (s_2, a_*) or (s_3, a_*) with a probability of 0.4, or in $(s_2, \overline{a})$ with a probability of 0.2. MCES is then expected to cross the boundary in state s_3 and transition to policy π_2.
2. When $q_k \in Q(\pi_2)$, start each episode in $(s_1, \overline{a})$. MCES will cross the boundary in s_1 and transition to π_3.
3. When $q_k \in Q(\pi_3)$, start each episode in (s_1, a_*) or (s_2, a_*) with a probability of 0.4, or in $(s_1, \overline{a})$ with a probability of 0.2. MCES is then expected to cross the boundary in s_2 and transition to π_4.
4. When $q_k \in Q(\pi_4)$, start each episode in $(s_3, \overline{a})$. MCES will cross the boundary in s_3 and transition to π_5.
5. When $q_k \in Q(\pi_5)$, start each episode in (s_1, a_*) or (s_3, a_*) with a probability of 0.4, or in $(s_3, \overline{a})$ with a probability of 0.2. MCES is then expected to cross the boundary in s_1 and transition to π_6.
6. When $q_k \in Q(\pi_6)$, start each episode in $(s_2, \overline{a})$. MCES will cross the boundary in s_2 and transition to π_1.

Sampling the initial state-action pair can interrupt the illustrated sequence. For example, updates to $(s_2, \bar{a})$ during step 1 can take the policy to π_6, after which step 6 returns it to π_1. Nevertheless, the sampling strategy guarantees that, for every initial condition in $\mathcal{Q}_0$, the greedy policy π_k contains at least one suboptimal action but never all three. The policy remains suboptimal, although it may eventually stop changing. This behaviour does not rely on the exact probabilities above; it depends only on what pairs are updated. It persists when increasing the sampling bias towards greedy actions, for instance when replacing the 2-to-1 sampling ratio of (s_2, a_*) to $(s_2, \overline{a})$ during step 1 with an ϵ-greedy strategy for small ϵ.

We incorporate strict exploring starts by initialising each episode using the sampling strategy above with a probability of 0.9 or by uniformly sampling a state-action pair with a probability of 0.1. Since the wrong state-action pairs are now sampled at the wrong time infinitely often, solutions can visit the policies π_* or $\overline{\pi}$. When they do, we sample the initial state-action pair uniformly. Additionally, we introduce stochastic transitions by terminating each episode with a probability of 0.1 after each action. The value $q_{\pi_k}(s, a)$ is then always over- or underestimated, which yields steps that potentially drive solutions out of $\mathcal{Q}_0$. Solutions with exploring starts cannot remain trapped in a sub-optimal policy.

Despite the variability of initial state-action pairs and the noisy action-value estimates, IV MCES is observed to converge to a suboptimal solution when using SA learning rates with high probability. For instance, 99.6% of solutions converge to a point where $q(s, a_*) = q(s, \bar{a}) \approx 1.615$ for all states when starting in $q_0(s_1, a_*) = q_0(s_2, a_*) = q_0(s_3, \bar{a}) = 2$ and $q_0(s_1, \bar{a}) = q_0(s_2, \bar{a}) = q_0(s_3, a_*) = 1$ with $n_0(s, a) = 1$ for all state-action pairs (Figure 2). The remaining 0.4% escape to q_{π_*}.

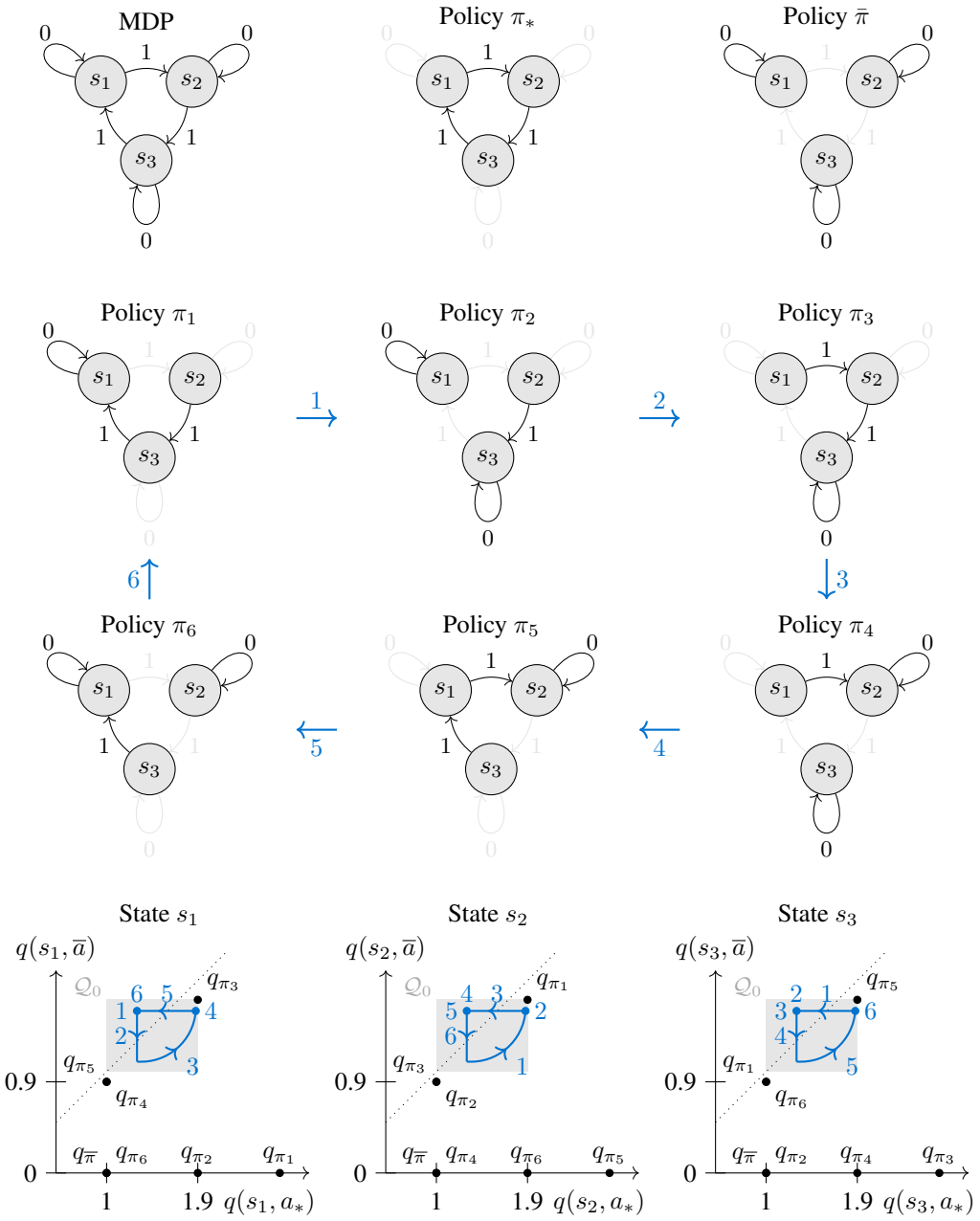


Figure 1: The sampling strategy described on page 3 ensures that IV MCES remains trapped in the set of six suboptimal policies for every initial condition in $\mathcal{Q}_0$ with one of these policies greedy. *Top:* the MDP with reward 1 for "move" (action a_*) and 0 for "stay" (action $\bar{a}$), the optimal policy π_* and the worst policy $\bar{\pi}$. *Middle:* the six suboptimal policies in the cycle. The blue arrow labelled i is the transition caused by step i of the sampling strategy. Each policy in the cycle differs from its predecessor in exactly one state. *Bottom:* the blue curve illustrates the intended trajectory of the estimates q_k, one panel per state s, in the coordinates $(q(s, a_*), q(s, \bar{a}))$. Each segment is labelled by the step of the sampling strategy that generates it. The dotted diagonal $q(s, a_*) = q(s, \bar{a})$ is the boundary between the two greedy choices in s: above it action $\bar{a}$ is greedy, below it a_* is greedy. For example, during step 1 the greedy policy is π_1: in s_1, $\bar{a}$ is greedy and no episode starts there, so $q_k(s_1, \cdot)$ is not updated and remains at the top-left corner of the cycle; in s_2, a_* is updated twice as often as $\bar{a}$, so $q_k(s_2, \cdot)$ moves towards q_{π_1} along an arc; in s_3, only a_* is updated, so $q_k(s_3, \cdot)$ moves horizontally to the left until it crosses the dotted line. At that moment the greedy action in s_3 switches from a_* to $\bar{a}$, the solution enters $Q(\pi_2)$, and step 2 of the strategy takes over. Solutions remain in this set of 6 policies without ever reaching π_* or $\bar{\pi}$.

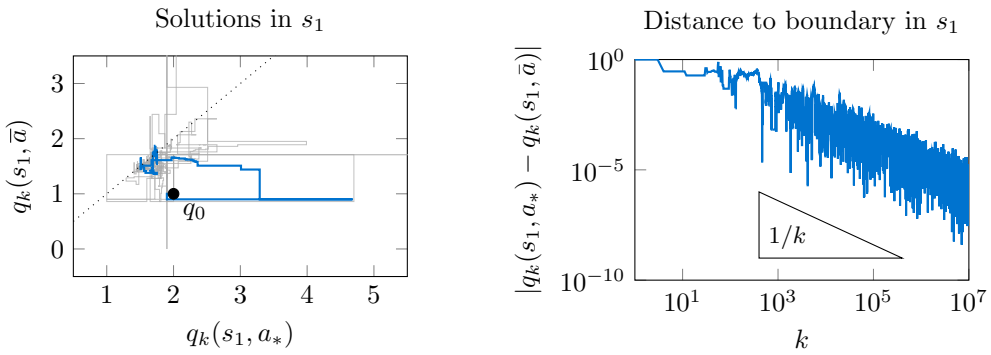


Figure 2: IV/SA MCES shows persistent suboptimal behaviour (representative solution in blue) even when all sampling distributions are bounded away from zero uniformly over episodes (strict exploring starts) and, in each state, give the greedy action no smaller probability than any other (greedy bias).

The mechanism which allows solutions to converge to the boundary despite frequently visiting the optimal region $Q(\pi_*)$ can be described as follows. Consider a solution in $Q(\pi_*)$ near the boundary point 1.615. So $q_k(s, a_*) > q_k(s, \bar{a})$ for every state, and $q_k(s, a) < q_{\pi_*}(s, a)$ for every state-action pair. There are two possible updates. Updating (s, a_*) moves q_k deeper into $Q(\pi_*)$ on average as $q_k(s, a_*)$ increases compared to $q_k(s, \bar{a})$. In contrast, updating $(s, \bar{a})$ guides the solution towards the boundary of $Q(\pi_*)$ as $q_k(s, \bar{a})$ catches up to $q_k(s, a_*)$. If the learning rate is large enough, $\bar{a}$ may even become greedy in s. Thus, solutions leave the optimal region if, on average, updates to $(s, \bar{a})$ receive larger learning rates or are more frequent than updates to (s, a_*). Precisely this behaviour is induced by the sampling strategy, the initial-visit updates and the sample-average learning rates. Specifically, the sampling strategy combined with initial-visit updates implies that, for every state, a_* is updated more frequently than $\bar{a}$ when solutions cycle between the policies π_1, π_2, π_3, π_4, π_5 and π_6. Therefore, $1/n_k(s, a_*)$ decreases faster than $1/n_k(s, \bar{a})$ for each state. As a result, when noise pushes solutions into the optimal region $Q(\pi_*)$, all actions are updated equally frequently but updates to $(s, \bar{a})$ are larger than updates to (s, a_*), which drives solutions out of $Q(\pi_*)$. In other words, the sampling strategy, the initial-visit updates and the learning rates effectively bias updates towards suboptimal actions when $\pi_k = \pi_*$ even though greedy actions are updated more frequently. This allows solutions to consistently jump in and out of the optimal region.

Overall, this example suggests that IV/SA MCES may converge to suboptimal solutions even when, in each state, the greedy action is updated at least as frequently as any other action. The asymptotic behaviour is determined by the interplay between learning rates and within-state update frequencies, rather than either factor alone. This suggests a remedy: scale learning rates to offset non-uniform update frequencies in each state and recover convergence to optimality.

3 Fix for initial-visit MCES

Starting from an initial q_0, IV MCES repeats the following three steps for increasing k. First, select a deterministic policy π_k such that $\pi_k(s) \in \arg\max_a q_k(s, a)$ for every state s. Then sample an initial state-action pair $(s_k, a_k) \sim \sigma_k$. Finally, update the estimate q_k as

$$q_{k+1}(s, a) = \begin{cases} q_k(s, a) + \eta_k(s, a)\,(\tilde{q}_k(s, a) - q_k(s, a)) & (s, a) = (s_k, a_k) \\ q_k(s, a) & \text{otherwise} \end{cases} \tag{5}$$

where $\tilde{q}_k$ is the return from the sampled episode under policy π_k, as defined in (2).

Section 2 and Appendix A show that IV MCES may behave suboptimally when the sampling distributions are biased towards greedy or non-greedy actions. This section presents a way to compensate for such non-uniform sampling by appropriately normalising the learning rates and thereby restoring convergence to optimality. The approach relies on the chain rule decomposition

$$\sigma_k(s, a) = f_k(s)\, g_k(a|s). \tag{6}$$

Here $f_k(s) := \sum_{a \in \mathcal{A}(s)} \sigma_k(s, a)$ is the probability that episode k starts in state s, and $g_k(a|s) := \sigma_k(s, a)/f_k(s)$ is the conditional probability that it starts with action a given that it starts in state s. When $f_k(s) = 0$, define $g_k(\cdot \mid s)$ as the uniform distribution over the available actions.

To formalise this normalisation, let $\mathcal{F}_k$ denote the information available after computing q_k, but before selecting policy π_k, sampling the initial pair (s_k, a_k), and generating the episode $\mathcal{E}_k$.

Theorem 1. *Assume*

(1) ***Discounted or episodic finite MDP****: the state and action sets are finite and rewards have finite second moments; and either $\gamma \in [0, 1)$, or $\gamma = 1$ and every deterministic stationary policy reaches a terminal state almost surely from every state.*

(2) ***Inertia of greedy policies****: when several actions are greedy, we retain the previous action wherever it is still greedy, meaning that*

$$\forall\, s \in \mathcal{S}: \quad \pi_k(s) \in \arg\max_{a \in \mathcal{A}(s)} q_{k+1}(s, a) \quad \Longrightarrow \quad \pi_{k+1}(s) = \pi_k(s). \tag{7}$$

(3) ***Infinite learning opportunities****: the sampling distributions $\{f_k\}_{k\geq 0}$ satisfy*

$$\forall\, s \in \mathcal{S}: \quad \sum_{k=0}^{\infty} \alpha_k f_k(s) = \infty \tag{8}$$

almost surely, and the sampling distributions $\{g_k\}_{k\geq 0}$ satisfy

$$\forall\, k \geq 0,\ \forall\, s \in \mathcal{S},\ \forall\, a \in \mathcal{A}(s): \quad g_k(a|s) \geq g_{\min} \tag{9}$$

for some constant $g_{\min} \in (0, 1)$.

(4) ***Robbins–Monro and weakly monotone learning rates****: the deterministic, scalar sequence $\{\alpha_k\}_{k\geq 0}$ satisfies*

$$\sum_{k=0}^{\infty} \alpha_k = \infty, \qquad \sum_{k=0}^{\infty} \alpha_k^2 < \infty, \tag{10}$$

and there exist a time $K_\alpha < \infty$ and a constant $c_\alpha \geq 1$ such that

$$\forall\, j \geq k \geq K_\alpha: \quad \alpha_j \leq c_\alpha \alpha_k. \tag{11}$$

Given that condition (10) *requires that $\alpha_k \to 0$ and we study the asymptotic behaviour of MCES, we assume, without loss of generality, that $\alpha_k \in (0, 1]$.*

Then IV MCES converges almost surely to q_{π_*} when using learning rates

$$\eta_k(s, a) := \frac{\alpha_k}{g_k(a|s)}. \tag{12}$$

The scaling in (12) is a local analogue of sample-average learning rates. While $1/n_k(s, a)$ gives a greater weight to actions that have been sampled less often in the past, the factor $1/g_k(a|s)$ gives a greater weight to actions which are less likely to be sampled at the current step. The correction is thus local in time: it compensates for the current sampling imbalance rather than the historical one.

Theorem 1 is an adaptation of the main convergence result in [13] because the normalisation in (12) equalises the rate of change across actions on a state-by-state basis at every step. As a result, the associated mean-field dynamics generate monotonically improving policies. This structure then allows us to establish convergence of the original stochastic solutions by applying an extension of the combined stability-ODE method [14]. We sketch the proof below, and provide further details on the adaptation of [13] in Appendix B.

Proof sketch. Equation (5) with learning rates as in (12) is equivalent to

$$\begin{aligned} q_{k+1}(s, a) &= q_k(s, a) + \alpha_k \left(\frac{\sigma_k(s, a)}{g_k(a|s)} \big(q_{\pi_k}(s, a) - q_k(s, a) \big) + w_k(s, a) \right) \\ &= q_k(s, a) + \alpha_k \left(f_k(s) \big(q_{\pi_k}(s, a) - q_k(s, a) \big) + w_k(s, a) \right) \end{aligned} \tag{13}$$

for every state-action pair. Here, w_k incorporates the perturbations caused by the episode simulation and the asynchronous updates of state-action pairs. Given that $\tilde{q}_k$ is an unbiased estimate of q_{π_k} under initial-visit updates, it follows that $\mathbb{E}[w_k(s, a) \mid \mathcal{F}_k] = 0$.

Consider the mean-field dynamics underlying (13):

$$q_{k+1}(s,a) = q_k(s,a) + \alpha_k f_k(s)(q_{\pi_k}(s,a) - q_k(s,a)). \tag{14}$$

Since the new policy π_{k+1} will be greedy with respect to q_{k+1}, we have $q_{k+1}(s,\pi_{k+1}(s)) \geq q_{k+1}(s,\pi_k(s))$ for every state s. Using (14), this becomes

$$\begin{aligned} q_k(s,\pi_{k+1}(s)) + \alpha_k f_k(s)\big(q_{\pi_k}(s,\pi_{k+1}(s)) - q_k(s,\pi_{k+1}(s))\big) \\ \geq q_k(s,\pi_k(s)) + \alpha_k f_k(s)\big(q_{\pi_k}(s,\pi_k(s)) - q_k(s,\pi_k(s))\big). \end{aligned}$$

Rearranging the inequality yields

$$q_{\pi_k}(s,\pi_{k+1}(s)) - q_{\pi_k}(s,\pi_k(s)) \geq \frac{1-\alpha_k f_k(s)}{\alpha_k f_k(s)}\big(q_k(s,\pi_k(s)) - q_k(s,\pi_{k+1}(s))\big).$$

If $f_k(s) = 0$, then state s was not updated. So condition (7) implies that $\pi_{k+1}(s) = \pi_k(s)$, and consequently $q_{\pi_k}(s,\pi_{k+1}(s)) = q_{\pi_k}(s,\pi_k(s))$. Alternatively, if $f_k(s) > 0$, it follows that

$$q_{\pi_k}(s,\pi_{k+1}(s)) \geq q_{\pi_k}(s,\pi_k(s))$$

because π_k is greedy with respect to q_k and $0 < \alpha_k f_k(s) \leq 1$. As a result, the Policy Improvement Theorem guarantees that the new policy π_{k+1} is at least as good as policy π_k [4].

Finally, we enforce improving transitions in (13) by maintaining a positive distance to "bad" regions using two events with fixed non-zero probability. A *seed event* first forces the distance to $O(\alpha_k)$, with probability $p_s > 0$. A *growth event* then keeps it strictly positive until the next transition, with probability $p_g > 0$. Since, at any time, the next policy will be better with probability at least $p_s p_g > 0$, the probability of never reaching and remaining in the optimal region decays as $\big(1-(p_s p_g)^{|\mathcal{P}|}\big)^n$, where $\mathcal{P}$ is the set of deterministic policies. This probability tends to zero as n tends to infinity because the number of deterministic policies is finite. As a result, the complement event, i.e. locking into the optimal region and $q_k \to q_{\pi_*}$, occurs eventually with probability 1. ■

The assumptions in Theorem 1 are sufficient conditions that allow the probabilistic argument in the proof. However, some can likely be relaxed because they are not necessary for the core monotone-improvement mechanism: the inertia of greedy policies is used only to remove boundary ambiguity, which rarely occurs in practice; the uniform lower bound on g_k controls the factor $1/g_k$; and the weakly monotone learning rates control the potential adverse effects of noise over a finite number of steps. In contrast, a discounted or episodic finite MDP is necessary to apply the Policy Improvement Theorem; the infinite accumulated state sampling, which implies the first Robbins–Monro condition in (10), guarantees that updates are large enough to avoid premature convergence due to Zeno-like behaviour; and the second Robbins–Monro condition gradually forces the learning rates to zero so that the accumulated stochastic perturbations cannot prevent convergence.

Since $g_k(\cdot|s)$ is a distribution over $\mathcal{A}(s)$, its values scale as $1/|\mathcal{A}(s)|$. Hence, $1/g_k(a|s)$ is typically larger in states with more available actions. If necessary, dividing the learning rates in (12) by $|\mathcal{A}(s)|$ cancels out this effect. Since this additional factor is constant across actions within a state, it does not affect the argument. It only introduces extra state-dependent scaling factors in the proof.

Earlier work recovers convergence to optimality by scaling the learning rates with $1/\sigma_k$ [7, 9], requiring knowledge of the sampling distribution over *all* state-action pairs. Our normalisation uses only the conditional initial-action probabilities g_k, which is more practical for real-world applications with large state spaces. For instance, chess has $O(10^{45})$ different board layouts but only $O(10^1)$ legal moves per layout. So keeping track of a distribution over the joint state-action space is not feasible, whereas representing the action distribution within a given state is straightforward. Our normalisation also naturally recovers the more general convergence result in [13] because g_k is often known in practice, for example when episodes are initialised using an ϵ-greedy policy. Yet this normalisation is neither necessary nor optimal; depending on the structure of the MDP or the initialisation of MCES, convergence to optimality can still hold under non-uniform updates and sometimes even be faster. If g_k is unknown, estimating it is a plausible extension but the robustness of convergence to estimation errors on g_k strongly depends on the parameters of the MDP, as illustrated in Appendix A.

The earlier work is based on the theorem of Tsitsiklis in [7], which uses commutativity in a central way. Since this technique does not work when updates are only uniform over the actions within each state, a different mathematical approach is necessary to prove Theorem 1.

4 Counterexamples for first-visit MCES with sample-average updates

4.1 A thirteen-state example with evidential support

Under first-visit updates, the state-action pairs updated after an episode depend on the transition dynamics induced by the greedy policy on the MDP. FV MCES is thus harder to analyse than IV MCES and counterexamples are harder to construct. This section presents such a counterexample: a cyclic MDP with 13 states generalising Figure 1, on which FV/SA MCES frequently converges to a suboptimal solution. Each state has two actions: a_* gives reward 1 and moves from s_i to s_{i+1} with probability τ or terminates the episode with probability $1-\tau$; action $\bar{a}$ yields no reward and stays in the same state.

We found the example by first analysing FV/SA MCES without noise, namely when $q_k(s,a)$ is updated towards the exact value $q_{\pi_k}(s,a)$, and $n_k(s,a)$ is incremented by the probability of updating (s,a) after an episode simulated with policy π_k. For $N=3$ this deterministic process does not converge to a suboptimal solution. For $N=4$ it can oscillate indefinitely between suboptimal policies, but this behaviour is fragile as it requires that the greedy policy changes in two states simultaneously. For $N \geq 5$ the same type of oscillation can proceed through one-state policy changes; the oscillation alternates between policies with one and two "stay" actions, as depicted in Figure 3. This suboptimal behaviour persists under the noise of the fully sampled FV/SA MCES process for sufficiently late starts, becoming increasingly robust to fluctuations early in the simulation as N grows. We describe below the case $N=13$.

The deterministic solutions oscillate indefinitely between the 26 policies when initialising episodes as follows. At each time step, with probability ϵ_{ES} sample the initial state-action pair uniformly amongst the 26 pairs; otherwise, choose it according to the current greedy policy. If that policy has a single "stay" action at s_i, start the episode in $(s_{i-6}, \bar{a})$. If the policy has adjacent "stay" actions at s_i and s_{i-1}, start in $(s_{i+1}, \bar{a})$ with probability λ or in (s_{i-1}, a_*) with probability $(1-\lambda)$. All state indices are to be taken modulo 13. Under this initialisation scheme, Figure 4 shows that the deterministic solutions spiral into a fixed point on the boundary. This behaviour is robust as a large set of initial conditions produces this spiral. We derive the value $\widehat{q}$ in Appendix C.2, and show that the subsequence $\{q_{k+26i}\}_{i\geq 0}$ moves along a straight line towards that fixed point in Appendix C.5. As a result, it is guaranteed that the deterministic solutions visit only these 26 policies.

Fully sampled FV/SA MCES solutions are extremely noisy, due to both sample returns and exploring starts. As a result, they frequently visit off-cycle policies. We then initialise episodes to guide the process back to an on-cycle policy as follows. On visiting the optimal all-"move" policy, we start episodes in the "stay" action of the state for which the action-values are closest to the boundary. The first visit structure ensures that "move" is then also updated in that, and potentially downstream, states,

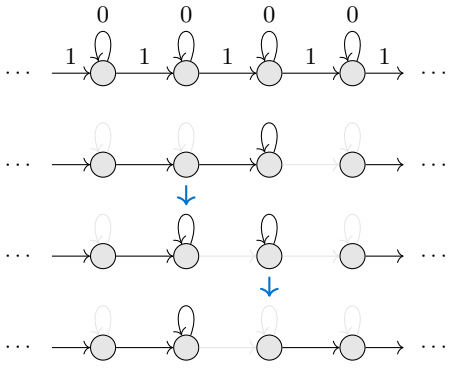


Figure 3: The MDP forms a connected chain of 13 states, each with two actions. As in Figure 1, solutions cycle between suboptimal policies that only contain 1 or 2 suboptimal actions. The suboptimal action propagates backwards relative to the state-transitions.

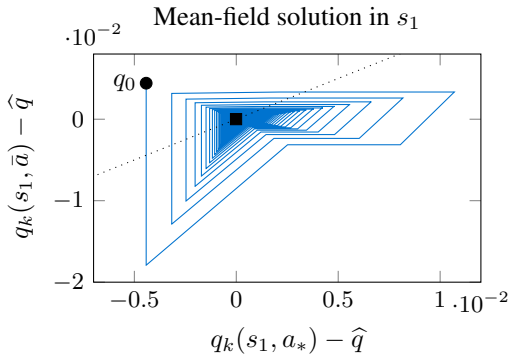


Figure 4: Mean-field solutions of FV/SA MCES robustly converge to a suboptimal point $(\widehat{q}, \widehat{q})$ derived in Appendix C.2. The solution is initialised with counts $n_0(s,a_*) := 10$ and $n_0(s,\bar{a}) := 5$. The dotted line corresponds to the boundary where $q(s_1,a_*) = q(s_1,\bar{a})$.

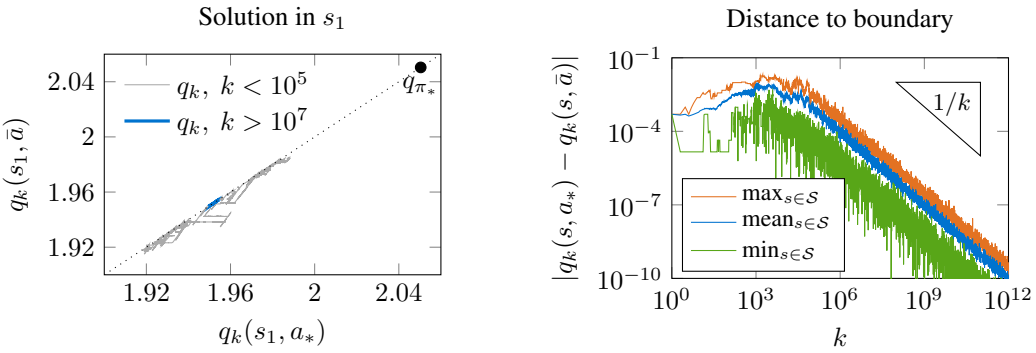


Figure 5: FV/SA MCES shows persistent suboptimal behaviour despite strict exploring starts.

and also that "move" actions are updated more often than "stay" actions overall. The $1/n_k(s, a)$ weighting then results in solutions being pushed towards and eventually across the boundary, creating a "stay" state. In all other off-cycle policies, we first consider all states where "stay" is greedy and keep only those with the fewest neighbours that also prefer "stay". Among that subset, we choose the state whose action values are closest to the decision boundary, then initialise the episode in the "stay" action of that state. This eventually results in solutions crossing the boundary in the opposite direction, reducing the number of "stay" states until only a singleton or an adjacent pair remains. Exploring starts remain active during the recovery. The sampling strategy is thus completely determined by the current q_k (as would be the case for an ϵ-greedy policy, for example) with a uniform lower bound on the probability of sampling any state-action pair at each step.

We simulated 40,000 solutions of FV/SA MCES with $\gamma = 0.9999$, $\tau \approx 0.5123$ and $\epsilon_{\text{ES}} = 10^{-4}$ for 10^9 steps for a range of initial counts $n_0(s, \bar{a}) = M$, $n_0(s, a_*) = 2.01M$ with M from 3,000 to 12,000 and found that all but a proportion of approximately $(2{,}200/M)^{5.4}$ remained trapped in a shrinking strip around the boundary significantly below the optimal value. In each run, for approximately 14.9% of steps the greedy policy was the all-"move" policy π_* and for approximately 1.5% of steps it was another off-cycle policy. Thus for a large majority of steps the system is in a cycle phase. For a representative sample of runs which had not escaped at 10^9 steps we continued the simulation up to 10^{12} steps and none escaped. We then ran 8 simulations from $M = 10^5$ and all remained trapped at 10^{12} steps. That is, all trials that survived 10^9 steps went on to survive 10^{12} steps. Figure 5 shows representative behaviour from a start at $M = 1{,}000$, demonstrating an initial lock-in followed by the long $1/k$ tail. Altogether, these simulations provide strong evidence of high-probability non-convergence to optimality from late starts of FV/SA MCES for this MDP.

This counterexample is intuitive, but hard to study analytically. The two-state example in the next section is perhaps less intuitive, but simple enough to admit a non-convergence theorem.

4.2 A two-state example with analytical proof

Consider an episodic two-state MDP with $\gamma = 1$ and actions a_* and $\bar{a}$ in each state. At s_1, action a_* yields no reward and moves to s_2 with probability $1/3$ and otherwise terminates; at s_2, it returns to s_1 with reward zero. At either state, $\bar{a}$ self-loops with zero reward with probability $1 - \tau$, and otherwise terminates with loss 1 at s_1 or $L = 26/7$ at s_2. Every stationary policy terminates almost surely. The optimal policy π_* chooses a_* everywhere, with $q_{\pi_*} = (0, -\tau, 0, -L\tau)$, where state-action pairs are ordered as $((s_1, a_*), (s_1, \bar{a}), (s_2, a_*), (s_2, \bar{a}))$. For this example, $\mathcal{P} = \{\pi_*, \pi_1, \pi_2, \bar{\pi}\}$, where π_i chooses a_* only in s_i $(i = 1, 2)$ and $\bar{\pi}$ chooses $\bar{a}$ everywhere. Choose starts uniformly with probability ϵ_{ES}, otherwise choose (s_2, a_*) in π_1, (s_1, a_*) in π_2, $(s_2, \bar{a})$ in $\bar{\pi}$; and in π_* choose $(s_1, \bar{a})$ with probability p_{s} and $(s_2, \bar{a})$ otherwise, with $p_{\text{s}} = 10/51$. Thus each pair has a start probability of at least $\epsilon_{\text{ES}}/4$. Resolve greedy ties in favour of a_*, and initialise the counts as $n_0(s, a) = 0$ for every state-action pair.

Theorem 2. *Let $0 < \tau \le 10^{-8}$ and $0 < \epsilon_{\text{ES}} \le 4 \times 10^{-11}$, and run FV/SA MCES as above on the two-state MDP from $q_0 = (-L, 0, -L, -L)$. Let $q^\dagger := (-1/6, -1/6, -1/5, -1/5)$. Then there is a time K such that, with positive probability, for all $k \ge K$*

$$\|q_k - q^\dagger\|_\infty < 5 \times 10^{-6}, \qquad \|q_k - q_{\pi_*}\|_\infty > 1/5 - 5 \times 10^{-6}.$$

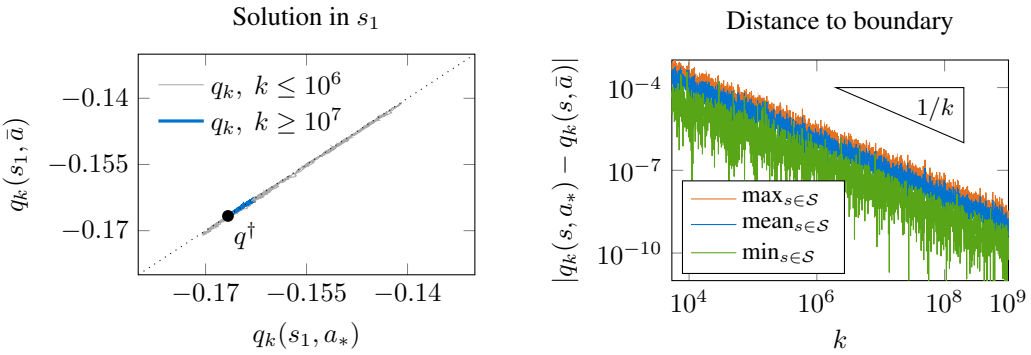


Figure 6: FV/SA MCES starting at $q_0 = q^\dagger$ with $\tau = 10^{-4}$ and $\epsilon_{\text{ES}} = 10^{-4}$.

The proof in Appendix D uses the exact sample-average identities to bound errors in the estimates and normalized counts in terms of the two action gaps and the frequency of policy $\bar{\pi}$. We further bound this frequency using the number of episodes started from the uniform exploring start component. Probability bounds then show that the estimates remain near $q^\dagger$ for all subsequent episodes. Reaching the late start from an empty initialization is a straightforward finite construction, as explained at the end of the appendix. Thus strict exploring starts do not guarantee optimal convergence for this class. The small parameters and large sufficient counts are conservative; Figure 6 shows a finite sampled run starting from an earlier start with estimates $q^\dagger$ and with more realistic parameters.

5 Conclusion

We have presented initial-visit and first-visit examples of persistent suboptimal behaviour, including a theorem for first-visit MCES on a two-state MDP, and proposed a practical fix that guarantees convergence to optimality in the initial-visit case.

Previous counterexamples for initial-visit MCES rely on sampling distributions that heavily favour non-greedy actions for every episode start. Our example shows that stable suboptimal solutions exist even when greedy actions are updated more frequently, as typically occurs during the exploitation phase. The suboptimal behaviour arose because the sample-average learning rates introduced an effective bias towards non-greedy actions when solutions were in the optimal region.

We then presented a practical learning-rate scheme that restores convergence to optimality despite updating actions at different frequencies. The key idea is to scale learning rates inversely with update frequencies on a state-by-state basis, so that the effective update size is uniform across actions within each state. Under this correction, the underlying mean-field generates monotonically improving policies. Since the stochastic perturbations cannot consistently prevent this improvement, initial-visit MCES eventually converges to the optimal policy and its action values.

Finally, we have shown that first-visit MCES with sample-average updates can also have stable suboptimal solutions despite strict exploring starts. On a 13-state MDP, simulations showed that sampled solutions lock into a stable suboptimal attractor of the mean-field dynamics. On a two-state MDP, we proved that, with strictly positive probability, solutions remain trapped at a fixed distance from q_{π_*} forever. This settles the open problem posed by Sutton and Barto [2, 4] in the negative: exploring starts alone do not guarantee convergence of FV/SA MCES to optimality.

Overall, these results highlight that the asymptotic behaviour of MCES depends critically on the interplay between the learning rates and the sampling distributions for initial state-action pairs of episodes. Together, these determine the effective update size for each action. Under the stated assumptions, our initial-visit convergence result shows that making these effective updates uniform across actions within each state restores convergence to optimality. Characterising more general update schemes that retain this guarantee remains an important direction for future work.

Acknowledgements and Disclosures

This research was supported by the MathWorks Studentship.

References

[1] Satinder P. Singh and Richard S. Sutton. Reinforcement Learning with Replacing Eligibility Traces. *Machine Learning*, 22:123–158, 1996.

[2] Richard S. Sutton. Open Theoretical Questions in Reinforcement Learning. In *Proceedings of the 4th European Conference on Computational Learning Theory*, pages 11–17. Springer-Verlag, 1999.

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

[4] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. The MIT Press, 2nd edition, 2018.

[5] Dimitri P. Bertsekas and John N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, Massachusetts, 1996.

[6] Che Wang, Shuhan Yuan, Kai Shao, and Keith Ross. On the Convergence of the Monte Carlo Exploring Starts Algorithm for Reinforcement Learning. In *The Tenth International Conference on Learning Representations*, 2022.

[7] John N. Tsitsiklis. On the Convergence of Optimistic Policy Iteration. *Journal of Machine Learning Research*, 3:59–72, 2002.

[8] Yuanlong Chen. On the convergence of optimistic policy iteration for stochastic shortest path problem. Technical report, 2018.

[9] Jun Liu. On the convergence of reinforcement learning with Monte Carlo Exploring Starts. *Automatica*, 129:109693, 2021.

[10] Joseph Lubars, Anna Winnicki, Michael Livesay, and Rayadurgam Srikant. Optimistic Policy Iteration for MDPs with Acyclic Transient State Structure. Technical report, 2021.

[11] Rémi Munos, Thomas Stepleton, Anna Harutyunyan, and Marc G Bellemare. Safe and Efficient Off-Policy Reinforcement Learning. *Advances in Neural Information Processing Systems*, 29, 2016.

[12] Anna Winnicki and Rayadurgam Srikant. On The Convergence Of Policy Iteration-Based Reinforcement Learning With Monte Carlo Policy Evaluation. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, 2023.

[13] Octave Oliviers and Glenn Vinnicombe. Convergence of Monte Carlo Optimistic Policy Iteration: Beyond Uniform State-Action Updates. *arXiv:2606.10580*, 2026.

[14] Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer-Verlag, 2003.

[15] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

A Counterexample for initial-visit MCES with non-greedy bias

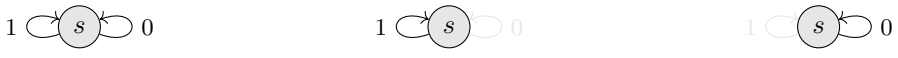
Bertsekas and Tsitsiklis show in [5, Example 5.12] that IV MCES can converge to suboptimal solutions when greedy actions are updated less frequently than non-greedy actions. The single-state MDP in Figure 7 is the simplest such example.

Episodes are initialised to favour the non-greedy action, meaning when $q_k(s, \pi_*(s)) \geq q_k(s, \bar{\pi}(s))$ the initial state-action pair is $(s, \pi_*(s))$ with a small probability ϵ and $(s, \bar{\pi}(s))$ with probability $1 - \epsilon$, with the probabilities reversed otherwise.

Figure 8a shows simulations with $\epsilon = 0.1$. The streamlines indicate the mean-field flow. Solutions starting to the right of the separatrix typically converge to q_{π_*}, whereas those starting to the left leave the optimal region $Q(\pi_*)$ and converge to a boundary point.

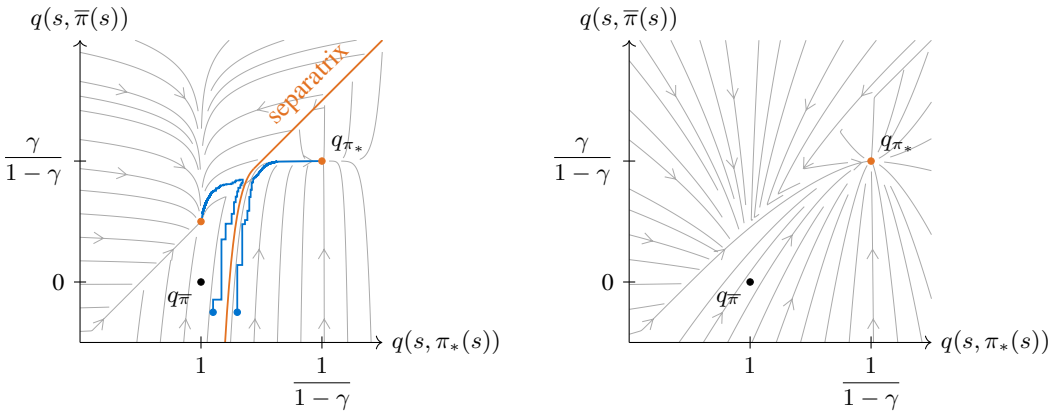
As ϵ increases to 0.5, the bias becomes more uniform and the mean-field flows in both policy regions progressively become straight lines towards their respective attractors. For some ϵ_* that satisfies $0 < \epsilon_* < 0.5$ and depends on the discount γ, a bifurcation occurs: the separatrix disappears and all mean-field solutions converge to q_{π_*}.

The normalisation $1/g_k(a|s)$ introduced in Theorem 1 effectively forces the updates in each state to be uniform, as if $\epsilon = 0.5$. If g_k is unknown, estimating it is a plausible practical extension. As long as the errors in g_k cause the effective ϵ to lie between ϵ_* and 0.5, MCES solutions will likely converge to q_{π_*}. Even though this claim remains unproven because it is difficult to rigorously link the stochastic solutions to mean-field ones, this example shows that robustness requires a case-by-case analysis.



(a) MDP (b) Policy π_* (c) Policy $\bar{\pi}$

Figure 7: MDP that exhibits a stable suboptimal solution for IV MCES when updates are biased towards non-greedy actions.



(a) Large non-greedy bias ($\epsilon = 0.1$) (b) Small non-greedy bias ($\epsilon = 0.4$)

Figure 8: The streamlines of the mean-field flow indicate that there are two basins of attraction when the non-greedy bias is large ($\epsilon = 0.1$). Solutions starting to the left of the separatrix tend to converge to a suboptimal fixed point on the boundary when the learning rates are sufficiently small. Those starting to the right converge to q_{π_*}. The two simulations of MCES in blue confirm this conclusion. As ϵ increases to 0.5, the bias becomes more uniform, the separatrix disappears and all mean-field solutions converge to q_{π_*}.

B Proof of Theorem 1

Let $\mathcal{F}'_k$ denote the information available after selecting the policy π_k and sampling distribution σ_k, but before sampling the initial state-action pair and generating its episode. Thus $\mathcal{F}_k \subseteq \mathcal{F}'_k$. Define the indicator $u_k(s,a) := \mathbf{1}_{\{(s_k,a_k)=(s,a)\}}$, and let $v_k(s,a) := \widetilde{q}_k(s,a) - q_{\pi_k}(s,a)$ at the updated pair, setting $v_k(s,a) = 0$ elsewhere. Combine these two sources of noise into a single stochastic perturbation w_k defined as

$$w_k(s,a) := \frac{1}{g_k(a|s)}\Big(u_k(s,a)v_k(s,a) + \big(u_k(s,a) - \sigma_k(s,a)\big)\big(q_{\pi_k}(s,a) - q_k(s,a)\big)\Big). \tag{15}$$

The uniform lower bound on g_k defined in (9) ensures that w_k is well defined. Equation (5) with learning rates as in (12) is then equivalent, for every state-action pair, to

$$\begin{aligned} q_{k+1}(s,a) &= q_k(s,a) + \alpha_k \frac{u_k(s,a)}{g_k(a|s)}\big(q_{\pi_k}(s,a) + v_k(s,a) - q_k(s,a)\big) \\ &= q_k(s,a) + \alpha_k\big(f_k(s)(q_{\pi_k}(s,a) - q_k(s,a)) + w_k(s,a)\big) \end{aligned} \tag{16}$$

where π_k is a deterministic policy that is greedy with respect to q_k.

We adapt the main convergence proof of [13]. Equation (16) has the same within-state drift, with $f_k(s)$ replacing the per-action probability $\sigma_k(s)$ used there. Since $g_{\min} \le g_k(a|s) \le 1$, the normalisation preserves the required summability and weak-monotonicity bounds. Moreover, $\alpha_k/g_k(a|s) \le 1$ for all sufficiently large k, so the fixed-target averaging and boundedness arguments of [13] carry over. The following lemma supplies the corresponding state-proportional noise bound.

Lemma 1. *There exists a constant $c_w \in [0,\infty)$ such that, for every $k \ge 0$, $s \in \mathcal{S}$, and $a \in \mathcal{A}(s)$,*

$$\mathbb{E}[w_k(s,a) \mid \mathcal{F}'_k] = 0, \qquad \mathbb{E}[w_k^2(s,a) \mid \mathcal{F}'_k] \le c_w f_k(s)(1 + \|q_k\|^2). \tag{17}$$

Here $\|\cdot\|$ denotes the supremum norm.

Proof. Fix a time $k \ge 0$, a state $s \in \mathcal{S}$ and an action $a \in \mathcal{A}(s)$. Under initial-visit updates and condition (1) of Theorem 1, $\tilde{q}_k$ is an unbiased estimate of q_{π_k} with uniformly bounded variance [1]. Thus there exists a constant $c_v \in [0,\infty)$ such that

$$\mathbb{E}[v_k(s,a) \mid \mathcal{F}'_k, u_k(s,a) = 1] = 0, \qquad \mathbb{E}\big[v_k(s,a)^2 \mid \mathcal{F}'_k, u_k(s,a) = 1\big] \le c_v, \tag{18}$$

for every $k \ge 0$, wherever $\sigma_k(s,a) > 0$. Consequently,

$$\mathbb{E}\big[u_k(s,a)v_k(s,a) \mid \mathcal{F}'_k\big] = \sigma_k(s,a)\,\mathbb{E}\big[v_k(s,a) \mid \mathcal{F}'_k, u_k(s,a) = 1\big] = 0. \tag{19}$$

When $\sigma_k(s,a) = 0$, equation (19) holds directly because $u_k(s,a) = 0$ almost surely.

Also, $\mathbb{E}[u_k(s,a) \mid \mathcal{F}'_k] = \sigma_k(s,a)$, and both $\sigma_k(s,a)$ and $(q_{\pi_k}(s,a) - q_k(s,a))$ are $\mathcal{F}'_k$-measurable. It follows that

$$\begin{aligned} &\mathbb{E}\big[\big(u_k(s,a) - \sigma_k(s,a)\big)\big(q_{\pi_k}(s,a) - q_k(s,a)\big) \mid \mathcal{F}'_k\big] \\ &\qquad = \big(\mathbb{E}[u_k(s,a) \mid \mathcal{F}'_k] - \sigma_k(s,a)\big)\big(q_{\pi_k}(s,a) - q_k(s,a)\big) = 0. \end{aligned} \tag{20}$$

Since $g_k(a|s)$ is $\mathcal{F}'_k$-measurable and uniformly lower bounded, dividing (19) and (20) by $g_k(a|s)$, and summing the results yields

$$\mathbb{E}[w_k(s,a) \mid \mathcal{F}'_k] = 0. \tag{21}$$

For the second moment, the indicator property and (18) give

$$\begin{aligned} \mathbb{E}\big[u_k(s,a)v_k(s,a)^2 \mid \mathcal{F}'_k\big] &\le c_v\sigma_k(s,a), \\ \mathbb{E}\big[(u_k(s,a) - \sigma_k(s,a))^2 \mid \mathcal{F}'_k\big] &= \sigma_k(s,a)(1 - \sigma_k(s,a)). \end{aligned}$$

The cross term in $w_k^2(s,a)$ has zero conditional mean, because

$$u_k(s,a)v_k(s,a)\big(u_k(s,a) - \sigma_k(s,a)\big) = (1 - \sigma_k(s,a))u_k(s,a)v_k(s,a).$$

Expanding (15) therefore yields

$$\begin{aligned}\mathbb{E}[w_k(s,a)^2 \mid \mathcal{F}'_k] &\le \frac{\sigma_k(s,a)}{g_k(a \mid s)^2}\Big(c_v + (1-\sigma_k(s,a))\big(q_{\pi_k}(s,a) - q_k(s,a)\big)^2\Big) \\ &\le \frac{f_k(s)}{g_{\min}}\Big(c_v + \big(q_{\pi_k}(s,a) - q_k(s,a)\big)^2\Big) \\ &\le c_w f_k(s)(1+\|q_k\|^2), \end{aligned} \tag{22}$$

for a finite constant c_w, since the action values of the finitely many deterministic policies are uniformly bounded. This proves the lemma. ■

On the stopped, target-constant intervals used in the convergence proof of [13], the noise differences defining its bad-pair gaps retain the factor $f_k(s)$, by $(x-y)^2 \le 2x^2 + 2y^2$. Thus its gap-preservation argument carries over with $f_k(s)$ replacing $\sigma_k(s)$ and modified constants.

For the seeding argument of [13], each requested action has conditional probability at least $g_{\min}$, given the sampled state, and favourable normalised updates have increments no smaller than those with scalar step α_k because $\alpha_k/g_k(a|s) \ge \alpha_k$. The conditional-law construction from [13] thus carries over with modified constants, preserving the sampling and return laws at completed states. The positive-progress and repetition arguments in its main convergence proof therefore imply finitely many changes of the action-value target, eventual optimality of the selected policies, and

$$q_k \longrightarrow q_{\pi_*} \qquad \text{almost surely},$$

as claimed.

C 13-state first-visit MCES example

C.1 MDP, cycle, and notation

This appendix describes the discounted 13-state example with exploring starts. The discount and exploring-start perturbation level are

$$\gamma = 0.9999, \qquad \epsilon_{\mathrm{ES}} = 10^{-4}.$$

The "move" action is denoted by a_* and the "stay" action by $\bar{a}$. A successful "move" sends

$$s_i \longmapsto s_{i+1},$$

with indices understood modulo 13. See Figure 3. Action a_* gives reward 1 and succeeds with probability τ; on success it moves to the next state, and on failure the episode terminates. Action $\bar{a}$ gives reward 0 and remains in the same state.

A deterministic greedy policy is represented by its "stay" set $S \subseteq \{s_1, \dots, s_{13}\}$. Its value function satisfies

$$V_S(s_i) = \begin{cases} \gamma V_S(s_i), & s_i \in S, \\ 1 + \gamma\tau V_S(s_{i+1}), & s_i \notin S, \end{cases}$$

and the exact policy action values are

$$q_S(s_i, a_*) = 1 + \gamma\tau V_S(s_{i+1}), \qquad q_S(s_i, \bar{a}) = \gamma V_S(s_i).$$

Because $\gamma < 1$, "stay" states satisfy $V_S(s_i) = 0$.

The intended adjacent cycle is

$$\{s_1\}, \{s_1, s_{13}\}, \{s_{13}\}, \{s_{13}, s_{12}\}, \dots, \{s_2\}, \{s_2, s_1\}. \tag{23}$$

Equivalently, the local pattern is

$$\{s_i\} \longrightarrow \{s_i, s_{i-1}\} \longrightarrow \{s_{i-1}\}.$$

The normal start laws will be chosen later.

There are two requirements for a deterministic cycle. First, it must converge to a point on the boundary; second, it must follow the right set of greedy policies as it does so. The first condition we call the full-cycle balance condition and the second the sign-margin criterion. These are the subjects of the next two sections.

C.2 General full-cycle balance equations

We require that, provided the trajectory follows the right policies, it converges onto the boundary.

Let $\mathcal{Z} = \{(s_i, a) : 1 \le i \le 13,\ a \in \{a_*, \bar{a}\}\}$ be the set of state-action starts. For phase j in the intended cycle, let S_j be the "stay" set, $q_j = q_{S_j}$ its exact action-values, and ν_j its nominal start distribution on $\mathcal{Z}$. We impose a rotational symmetry, that $\nu_{\{s_i\}}(s_{i+k}, a)$ is a function of k only, so it is only necessary to specify one distribution over $\mathcal{Z}$. With exploring-starts the mixture is

$$\nu_j^{\epsilon_{\mathrm{ES}}} = (1 - \epsilon_{\mathrm{ES}})\nu_j + \epsilon_{\mathrm{ES}} U_{\mathcal{Z}}, \qquad U_{\mathcal{Z}}(z) = \frac{1}{26}.$$

Let $\mu_j^z(s, a)$ be the probability of visiting (s, a) when the episode starts from z and follows policy S_j after the start action. The phase-j first-visit probability under the perturbed start law is

$$\mu_j(s, a) = \sum_{z \in \mathcal{Z}} \nu_j^{\epsilon_{\mathrm{ES}}}(z) \mu_j^z(s, a).$$

For one complete pass through the 26 intended phases, define

$$B_a(s) = \sum_{j=1}^{26} \mu_j(s, a), \qquad A_a(s) = \sum_{j=1}^{26} \mu_j(s, a) q_j(s, a).$$

Here $B_a(s)$ is the expected number of first-visit updates to (s, a) per normal full cycle, and $A_a(s)$ is the corresponding expected return-weighted update mass.

If fractional expected updates are applied over one full cycle, then

$$N^+(s, a) = N(s, a) + B_a(s)$$

and

$$q^+(s, a) = \frac{N(s, a) q(s, a) + A_a(s)}{N(s, a) + B_a(s)}.$$

Equivalently,

$$q^+(s, a) - q(s, a) = \frac{B_a(s)}{N(s, a) + B_a(s)} \left(\frac{A_a(s)}{B_a(s)} - q(s, a) \right).$$

Thus the full-cycle target for coordinate (s, a) is $A_a(s)/B_a(s)$.

The boundary cancellation condition asks that, for every state,

$$\frac{A_{a_*}(s)}{B_{a_*}(s)} = \frac{A_{\bar{a}}(s)}{B_{\bar{a}}(s)} = \widehat{q}, \tag{24}$$

i.e. the target is on the boundary. By rotational symmetry it is enough to impose the scalar equation at one state,

$$F := A_{a_*}(s_1) B_{\bar{a}}(s_1) - A_{\bar{a}}(s_1) B_{a_*}(s_1) = 0. \tag{25}$$

The numerator of F is a high-order polynomial in τ and a low order polynomial in each component of ν.

The corresponding count ratio is

$$r = \frac{B_{a_*}(s)}{B_{\bar{a}}(s)}. \tag{26}$$

Choosing initial counts in this ratio makes the two action coordinates have matching first-order full-cycle learning rates:

$$\frac{B_{a_*}(s)}{N_0(s, a_*)} = \frac{B_{\bar{a}}(s)}{N_0(s, \bar{a})}.$$

C.3 Phase-sign margin LP

The balance equations center the deterministic cycle on the boundary, but they do not by themselves ensure that the intended greedy policies occur in the right order. Let

$$\sigma_j(s) = \begin{cases} +1, & s \in S_j, \\ -1, & s \notin S_j. \end{cases}$$

At the boundary level $\widehat{q}$, define the scaled expected gap increment in phase j by

$$g_j(s) = \mu_j(s, \bar{a})\big(q_j(s, \bar{a}) - \widehat{q}\big) - \frac{\mu_j(s, a_*)}{r}\big(q_j(s, a_*) - \widehat{q}\big). \tag{27}$$

Let

$$C_k(s) = \sum_{j<k} g_j(s)$$

be the cumulative gap displacement before phase k. The scaled initial gap $d_0(s)$ is then chosen by

$$\begin{aligned} \text{maximise} \quad & \rho, \\ \text{over} \quad & d_0 \in \mathbb{R}^{13},\ \rho \in \mathbb{R}, \\ \text{subject to} \quad & \sigma_k(s)\big(d_0(s) + C_k(s)\big) \geq \rho, \qquad k = 1, \ldots, 26, \quad s = s_1, \ldots, s_{13}. \end{aligned} \tag{28}$$

A positive optimal value ρ means that the phases are followed in the intended order.

The initial points are then placed symmetrically around the boundary:

$$q_0(s, a_*) = \widehat{q} - \frac{d_0(s)}{2N_0(s, \bar{a})}, \qquad q_0(s, \bar{a}) = \widehat{q} + \frac{d_0(s)}{2N_0(s, \bar{a})}.$$

The factor $1/2$ splits the desired total action gap equally above and below $\widehat{q}$.

C.4 Rotational start search and selected start law

The initial search for a suitable start distribution ν was over a wide class of rotationally symmetric distributions. Distributions with small support were preferred, as these would avoid adding too much extra noise into the full simulations. We started by considering all 13^2 pairs of deterministic starts. For each, we solved (25) for τ, and for any roots in $(0, 1]$ we then solved (28) for ρ. There were no solutions with a positive ρ.

We then broadened the search to distributions with a two-point support in the two-"stay" phases. That is, for offsets o_1, o_2, o_3 we set

$$\nu^{(o_1)}_{\{s_i\}} = \delta_{(s_{i+o_1}, \bar{a})}, \tag{29}$$

$$\nu^{(o_2, o_3, \lambda)}_{\{s_i, s_{i-1}\}} = \lambda\delta_{(s_{i+o_2}, \bar{a})} + (1-\lambda)\delta_{(s_{i+o_3}, a_*)}. \tag{30}$$

For a fixed triple and trial τ, equation (25) is solved for a root $\lambda \in [0, 1]$ and then the sign-margin LP (28) is evaluated. The τ that gives the maximum margin ρ is then found by gridding. The offsets $(o_1, o_2, o_3) = (-6, +1, -1)$ were chosen as giving a good margin with the corresponding λ close to 1.

Thus

$$\nu_{\{s_i\}} = \delta_{(s_{i-6}, \bar{a})}, \tag{31}$$

$$\nu_{\{s_i, s_{i-1}\}} = \lambda\delta_{(s_{i+1}, \bar{a})} + (1-\lambda)\delta_{(s_{i-1}, a_*)}, \tag{32}$$

and the corresponding τ and λ are given by

$$\tau = 0.512333524979486, \qquad \lambda = 0.923714016854384.$$

The resulting boundary quantities are

$$\widehat{q} = 1.971056244749326 \qquad q_* = \frac{1}{1-\gamma\tau} = 2.050366396036596 \qquad r = 2.010039006519156.$$

The sign-margin LP gives

$$\rho = 8.527647265071263 \times 10^{-3}$$

There were no distributions with a two point support in the single-"stay" phases and a one-point support in the two-"stay" phases which admitted a positive margin.

Note that this procedure is not specific to the 13-state MDP; it in fact gives a suitable three-point start distribution for the corresponding cycle on examples with 5 states and above, although a larger number of states gives a larger margin.

Figure 4 in the main text shows the resulting deterministic cycle. Note that the corners of the cycle contract towards $(\widehat{q}, \widehat{q})$ along a straight line, and so stay in the right greedy region; thus the cycles will persist indefinitely. The next section confirms this contraction.

C.5 Full-cycle linear interpolation toward $\widehat{q}$

Note that the full-cycle balance equations also imply a contraction. Suppose counts are exactly proportional to the full-cycle masses,

$$N_0(s, a_*) = c_s B_{a_*}(s), \qquad N_0(s, \bar{a}) = c_s B_{\bar{a}}(s).$$

Then after m complete normal cycles,

$$N_{26m}(s, a) = (c_s + m) B_a(s).$$

Using $A_a(s) = \widehat{q} B_a(s)$, the next full-cycle endpoint is

$$q_{26(m+1)}(s, a) = \frac{(c_s + m) B_a(s) q_{26m}(s, a) + A_a(s)}{(c_s + m + 1) B_a(s)} \tag{33}$$

$$= \frac{c_s + m}{c_s + m + 1} q_{26m}(s, a) + \frac{1}{c_s + m + 1} \widehat{q}. \tag{34}$$

Therefore the two-action point

$$P_m(s) = \big(q_{26m}(s, a_*), q_{26m}(s, \bar{a})\big)$$

satisfies

$$P_{m+1}(s) = (1 - \alpha_m) P_m(s) + \alpha_m (\widehat{q}, \widehat{q}), \qquad \alpha_m = \frac{1}{c_s + m + 1}. \tag{35}$$

Equivalently,

$$P_m(s) - (\widehat{q}, \widehat{q}) = \frac{c_s}{c_s + m} \big(P_0(s) - (\widehat{q}, \widehat{q})\big).$$

Thus the full-cycle endpoints move on the line segment toward the boundary point. The intermediate phase points need not lie precisely on this line, but deviate only by order $1/k$.

C.6 First-visit probabilities

Fix an anchor state s_i and write $d = (m - i) \bmod 13$, so $s_m = s_{i+d}$. Let μ_1^ε and μ_2^ε denote the perturbed first-visit probabilities in the one-"stay" and two-"stay" phases for the selected start law. The values for the current τ and λ are given in the following table for convenience.

Table 1: Single-phase first-visit probabilities by relative distance d for the tuned $N = 13$ start law.

d	$\mu_1^\varepsilon(d, \text{"move"})$	$\mu_1^\varepsilon(d, \text{"stay"})$	$\mu_2^\varepsilon(d, \text{"move"})$	$\mu_2^\varepsilon(d, \text{"stay"})$
0	$3.846154 \cdot 10^{-6}$	$1.809504 \cdot 10^{-2}$	$3.846154 \cdot 10^{-6}$	$3.908577 \cdot 10^{-2}$
1	$9.662821 \cdot 10^{-6}$	$3.846154 \cdot 10^{-6}$	$9.236313 \cdot 10^{-1}$	$9.236255 \cdot 10^{-1}$
2	$1.264289 \cdot 10^{-5}$	$3.846154 \cdot 10^{-6}$	$4.732150 \cdot 10^{-1}$	$3.846154 \cdot 10^{-6}$
3	$1.416969 \cdot 10^{-5}$	$3.846154 \cdot 10^{-6}$	$2.424516 \cdot 10^{-1}$	$3.846154 \cdot 10^{-6}$
4	$1.495191 \cdot 10^{-5}$	$3.846154 \cdot 10^{-6}$	$1.242238 \cdot 10^{-1}$	$3.846154 \cdot 10^{-6}$
5	$1.535267 \cdot 10^{-5}$	$3.846154 \cdot 10^{-6}$	$6.365169 \cdot 10^{-2}$	$3.846154 \cdot 10^{-6}$
6	$1.555800 \cdot 10^{-5}$	$3.846154 \cdot 10^{-6}$	$3.261859 \cdot 10^{-2}$	$3.846154 \cdot 10^{-6}$
7	$9.999157 \cdot 10^{-1}$	$9.999038 \cdot 10^{-1}$	$1.671929 \cdot 10^{-2}$	$3.846154 \cdot 10^{-6}$
8	$5.122980 \cdot 10^{-1}$	$3.846154 \cdot 10^{-6}$	$8.573545 \cdot 10^{-3}$	$3.846154 \cdot 10^{-6}$
9	$2.624751 \cdot 10^{-1}$	$3.846154 \cdot 10^{-6}$	$4.400207 \cdot 10^{-3}$	$3.846154 \cdot 10^{-6}$
10	$1.344825 \cdot 10^{-1}$	$3.846154 \cdot 10^{-6}$	$2.262066 \cdot 10^{-3}$	$3.846154 \cdot 10^{-6}$
11	$6.890759 \cdot 10^{-2}$	$3.846154 \cdot 10^{-6}$	$1.166624 \cdot 10^{-3}$	$3.846154 \cdot 10^{-6}$
12	$3.531136 \cdot 10^{-2}$	$3.846154 \cdot 10^{-6}$	$7.628220 \cdot 10^{-2}$	$6.015469 \cdot 10^{-4}$

C.7 Sampled recovery strategy

The balance equations and margin LP describe the intended normal cycle. In the sampled first-visit MC run, however, return noise, trajectory noise, and rare exploring starts can move the greedy policy off the intended 26-phase cycle. The simulation therefore has to choose start distributions even when the observed greedy "stay"-set is not one of the intended cycle phases.

Let

$$S = \{s : q(s, \bar{a}) > q(s, a_*)\}$$

denote the set of states for which the "stay" action is greedy at the start of an update. If S is one of the intended cycle phases, the controller uses the normal start law from Appendix C.4. If S is off-cycle, the rollout policy is the current greedy policy with "stay" set S, and the controller chooses from the 26 state-action starts as follows:

Let the current action gap be

$$G(s) = q(s, \bar{a}) - q(s, a_*).$$

If S is empty, then all $G(s)$ are negative. In this case we choose $s_0 = \arg\max_s G(s)$ as the state whose action values are nearest the "move"/"stay" boundary. For other non-cycle phases then for $s_i \in S$ we let $m(s_i) = \mathbf{1}(s_{i-1} \in S) + \mathbf{1}(s_{i+1} \in S)$ denote the number of neighbours of s_i in S and let $S_{\text{min}} = \{s \in S : m(s) = \min m\}$. That is, S_{min} is the subset of S consisting of the states with the minimum number of neighbours (all operations are modulo 13 and so s_1 and s_{13} are considered neighbours). Finally we let $s_0 = \arg\min_{s \in S_{\text{min}}} G(s)$ be the "stay" state with the fewest "stay" neighbours whose action values are closest to the "move"/"stay" boundary. This guarantees, for example, that if S consists of an adjacent pair and one or more isolated states then one of the isolated states is chosen, and in a run of adjacent states then the two ends are chosen in preference to an internal state. Finally, the starting pair is chosen as the "stay" action in this state $(s_0, \bar{a})$, in an attempt to move the action value across the boundary. If successful, this creates a "stay" state if S is empty and otherwise removes one, moving S closer to a cycle phase, otherwise. If at the next step S is still not a cycle phase then the process is repeated.

For the simulations we present, approximately 83.6% of updates are normal on-cycle updates, 14.9% are attempted recoveries from the optimal all-"move" policy and the remaining 1.5% are recoveries from all other non-cycle policies.

ChatGPT 5.5 was used to help refine and code this example. The core simulation was hand-coded.

D Proof for the two-state sampled example

The main step in the proof is showing a high probability of trapping from a late start. From the estimates and counts prescribed below, we keep the estimates and normalised counts close to fixed reference values at every later step. Exact sample-average identities reduce the problem to controlling the two action gaps and the use of the all-bad policy $\bar{\pi}$. We finish with a short comment on reaching this late start from an empty initialisation.

D.1 The two-state model and sample averages

This appendix proves the sampled-process result for a separate two-state example. The notation introduced here is local to this example. There are two nonterminal states, s_1, s_2, two actions, $a_*, \bar{a}$, and a terminal state ∂. At both states, we call a_* the good action and $\bar{a}$ the bad action. Returns are undiscounted: $\gamma = 1$. Put $L = 26/7$ and $p_{\mathrm{s}} = 10/51$. The transition and reward law is

pair	next state and reward	probability
(s_1, a_*)	$(s_2, 0),\ (\partial, 0)$	$1/3,\ 2/3$
(s_2, a_*)	$(s_1, 0)$	1
$(s_1, \bar{a})$	$(s_1, 0),\ (\partial, -1)$	$1-\tau,\ \tau$
$(s_2, \bar{a})$	$(s_2, 0),\ (\partial, -L)$	$1-\tau,\ \tau$

A key observation is that once either state is visited when the bad, self-looping, action is greedy at that state then termination with the stated negative reward is inevitable. τ determines the expected number of self loops, but not the eventual return. Throughout the proof,

$$0 < \tau \le 10^{-8}, \qquad 0 < \epsilon_{\mathrm{ES}} \le 4 \times 10^{-11}. \tag{36}$$

State-action pairs are ordered $((s_1, a_*), (s_1, \bar{a}), (s_2, a_*), (s_2, \bar{a}))$. We identify functions on these pairs with four-entry column vectors in this order. The policy set is $\mathcal{P} = \{\pi_*, \pi_1, \pi_2, \bar{\pi}\}$: π_* chooses a_* at both states, $\bar{\pi}$ chooses $\bar{a}$ at both, and π_r chooses a_* only at s_r, for $r = 1, 2$. Policy-indexed vectors and matrix columns use this order.

Before episode k, let $q_k(s, a)$ be the current return estimate for pair (s, a) and $n_k(s, a)$ its number of previously recorded first-visit returns. Choose $\pi_k \in \mathcal{P}$ greedily for q_k, resolving each tie in favour of a_*. Define four start distributions by

$$\begin{aligned} \bar{\sigma}^{(\pi_*)} &= (0, p_{\mathrm{s}}, 0, 1-p_{\mathrm{s}}), & \bar{\sigma}^{(\pi_1)} &= (0, 0, 1, 0), \\ \bar{\sigma}^{(\pi_2)} &= (1, 0, 0, 0), & \bar{\sigma}^{(\bar{\pi})} &= (0, 0, 0, 1), \\ \sigma^{(\pi)} &= (1-\epsilon_{\mathrm{ES}})\bar{\sigma}^{(\pi)} + \frac{\epsilon_{\mathrm{ES}}}{4}(1, 1, 1, 1). \end{aligned} \tag{37}$$

Sample the first pair from $\sigma^{(\pi_k)}$ and then follow π_k for the rest of the episode. Equivalently, choose a uniform start among the four pairs with probability ϵ_{ES}, and otherwise use $\bar{\sigma}^{(\pi_k)}$. Every pair therefore has start probability at least $\epsilon_{\mathrm{ES}}/4$. Thus the start distribution depends on the current greedy policy only.

Let $\mathcal{F}_k$ be all information available before episode $k \to k+1$, including q_k and the integer counts n_k. Conditional on this history, sample the initial pair and subsequent transitions according to the specified start distribution and the transition and reward laws above. Thus conditioning on $\mathcal{F}_k$ fixes everything already observed, while the next episode remains random. We write $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot \mid \mathcal{F}_k]$. Let $I_k(s, a)$ be one if pair (s, a) is visited in that episode and zero otherwise. Write I_k for its four-entry vector. Let $\tilde{q}_k(s, a)$ be the return from its first visit, set to zero when unvisited. All earlier rewards are zero, so the first-visit returns of all visited pairs equal the terminal reward. For each visited pair add one return to its sample average, so

$$\begin{aligned} n_{k+1}(s, a) &= n_k(s, a) + I_k(s, a), \\ q_{k+1}(s, a) &= q_k(s, a) + \frac{I_k(s, a)}{n_k(s, a) + 1}(\tilde{q}_k(s, a) - q_k(s, a)). \end{aligned} \tag{38}$$

An unvisited estimate is unchanged. Use the initialization $n_0 = 0$ and $q_0 = (-L, 0, -L, -L)$. Each initial value is used only to define the greedy policy until that pair is first visited, at which point it is overwritten by the observed return.

For a fixed policy π, let q_π be the vector of expected total returns from each initial pair, following π after the initial action.

Lemma 2. *Every fixed policy produces a finite episode with probability one and finite mean duration. Each visited pair's return is in $\{0, -1, -L\}$, and all visited pairs in one episode share its terminal reward. The unique optimal policy π_* chooses a_* in both states, with*

$$q_{\pi_*} = (0, -\tau, 0, -L\tau).$$

Proof. From any state and first action, absorption within the next two steps has probability at least τ: any initial bad action terminates with probability τ, a good action at s_1 with probability $2/3$, and a good action at s_2 reaches s_1, whose next action terminates with probability at least τ. The same bound applies after each surviving two-step block. Hence duration T satisfies $\mathbb{P}(T > 2m) \le (1-\tau)^m$, and so $\mathbb{E}[T] \le 2\sum_{m\ge 0}(1-\tau)^m = 2/\tau < \infty$. Only a terminal transition can earn non-zero reward, so all visited pairs share a return in the given set. No policy has positive return, whereas always choosing a_* gives zero. An initial bad action followed by this policy has expected return $-\tau$ at s_1 and $-L\tau$ at s_2. Both are strictly negative, proving the optimal vector and uniqueness of its greedy policy. ■

For $k \ge 1$ define

$$c_k(s,a) = \frac{n_k(s,a)}{k}, \qquad \widehat{c}_k(s,a) = \frac{n_k(s,a)+1}{k+1}. \tag{39}$$

Then $1/((k+1)\widehat{c}_k(s,a)) = 1/(n_k(s,a)+1)$ gives exactly the sample-average weight. Fix the reference values

$$q^\dagger = (-1/6, -1/6, -1/5, -1/5)^\top, \tag{40}$$

$$c^\dagger = (1, 3/16, 5/6, 65/96)^\top, \tag{41}$$

$$w^\dagger = (51/64, 7/64, 3/32, 0)^\top. \tag{42}$$

The coordinates of $q^\dagger, c^\dagger$ index state-action pairs; write $w^\dagger(\pi)$ for the component of $w^\dagger$ corresponding to policy π. The theorem below gives a trapping bound from value estimates $q^\dagger$ and prescribed normalized counts near $c^\dagger$. The vector $w^\dagger$ is a reference mixture of policies whose visit frequencies and return differences balance, as verified below; it does not describe the realized policy frequencies. Throughout this appendix, $\|v\|_\infty = \max_i |v_i|$, and the induced matrix norm is the maximum absolute row sum. The product $v \odot w$ is coordinatewise: $(v \odot w)_i = v_i w_i$. The unadorned vector $\mathbf{1}$ has four entries equal to one; for an event $\mathcal{E}$, the scalar indicator $\mathbf{1}_{\mathcal{E}}$ is one when $\mathcal{E}$ occurs and zero otherwise.

Theorem 3 (Conditional late-start trapping)**.** *Under* (36)*, let $k_0 \ge 10^{17}$ be any integer. Initialize* (38) *with $q_{k_0} = q^\dagger$ and positive integer counts satisfying $\|n_{k_0}/k_0 - c^\dagger\|_\infty \le 1/k_0$. There is an event on which, for every $k \ge k_0$,*

$$\|q_k - q^\dagger\|_\infty < 5 \times 10^{-6}, \qquad \|c_k - c^\dagger\|_\infty < 5 \times 10^{-6}. \tag{43}$$

On the same event, $\|q_k - q_{\pi_*}\|_\infty > 1/5 - 5 \times 10^{-6}$ for all $k \ge k_0$. Its probability conditional on the initial history is at least $1 - \beta_{\mathrm{trap}}(k_0)$, where $u = 2 \times 10^{-8}$, $\delta = 1.3 \times 10^{-7}$ and

$$\begin{aligned}\beta_{\mathrm{trap}}(k_0) = {} & 8e^{-2k_0u^2} + 8e^{-2k_0u^2/L^2} \\ & + 2e^{-k_0\delta^2/1058} + 2e^{-k_0\delta^2/200} + e^{-8(\log 2)10^{-11}k_0}.\end{aligned}$$

Floating-point evaluation gives $\beta_{\mathrm{trap}}(10^{17}) \approx 0.430$ and $\beta_{\mathrm{trap}}(10^{18}) \approx 2.31 \times 10^{-7}$.

D.2 A reference distribution for visits and returns

Suppress the episode index and write $I(s,a)$ and $\tilde{q}(s,a)$ for the visit indicators and first-visit returns. In policies $\pi_1, \pi_2, \bar{\pi}$, repeated bad actions change the episode length but not these quantities, so we need not count the repetitions.

The reference distribution uses the start law $\bar{\sigma}^{(\pi)}$ and in policy π_* requires the initial bad action to not terminate; the subsequent actions are then good and the return is zero. This comparison law is used only to compute fixed moments; the algorithm still uses the actual MDP with $\tau > 0$. The resulting finite distribution is given by the following table. Its last column gives $\tilde{q}(s,a)$ for each visited pair; unvisited pairs have return zero.

Class	Policy π	Probability conditional on π	Visit indicators I	$\tilde{q}(s,a)$ at visited pairs
A	π_*	$2p_{\mathrm{s}}/3$	$(1,1,0,0)$	0
B	π_*	$p_{\mathrm{s}}/3$	$(1,1,1,0)$	0
C	π_*	$1-p_{\mathrm{s}}$	$(1,0,1,1)$	0
F	π_1	$2/3$	$(1,0,1,0)$	0
G	π_1	$1/3$	$(1,0,1,1)$	$-L$
D	π_2	$2/3$	$(1,0,0,0)$	0
E	π_2	$1/3$	$(1,1,1,0)$	-1
—	$\bar{\pi}$	1	$(0,0,0,1)$	$-L$

For example, in class E the episode starts at (s_1, a_*), moves to s_2, returns to s_1 using a_*, and then repeats $\bar{a}$ at s_1 until it terminates with reward -1. In class C the initial action at $(s_2, \bar{a})$ survives and is followed by good actions. The first three visited pairs are $(s_2, \bar{a}), (s_2, a_*), (s_1, a_*)$; further good actions change neither the visit indicators nor the return.

Write $\mathbb{E}_\pi^{\mathrm{ref}}$ for expectation under the reference law for policy $\pi \in \mathcal{P}$, and define the fixed matrices

$$\begin{aligned} H_{(s,a),\pi} &= \mathbb{E}_\pi^{\mathrm{ref}} I(s,a), \\ F_{(s,a),\pi} &= \mathbb{E}_\pi^{\mathrm{ref}} [I(s,a)(\tilde{q}(s,a) - q^\dagger(s,a))]. \end{aligned}$$

Rows index pairs and columns index policies; $H_{\cdot,\pi}$ and $F_{\cdot,\pi}$ denote the columns labelled π. Thus H contains visit probabilities, and F contains expected return differences multiplied by visit indicators. The reference value $q^\dagger$ is fixed throughout the proof. Substituting $p_{\mathrm{s}} = 10/51$ and $L = 26/7$ in the table gives

$$H = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 10/51 & 0 & 1/3 & 0 \\ 133/153 & 1 & 1/3 & 0 \\ 41/51 & 1/3 & 0 & 1 \end{pmatrix}, \qquad F = \begin{pmatrix} 1/6 & -15/14 & -1/6 & 0 \\ 5/153 & 0 & -5/18 & 0 \\ 133/765 & -109/105 & -4/15 & 0 \\ 41/255 & -41/35 & 0 & -123/35 \end{pmatrix},$$

and so

$$Hw^\dagger = c^\dagger, \qquad Fw^\dagger = 0 \tag{44}$$

that is, if polices are selected in proportions $w^\dagger$ then state-action pairs are visited with relative frequency $c^\dagger$ and the average update at $q^\dagger$ is zero.

Lemma 3. *Put $\zeta = \tau + \epsilon_{\mathrm{ES}}$. Uniformly over histories, policies and pairs (s,a), the conditional expectations for actual episodes satisfy*

$$\begin{aligned} \left|\mathbb{E}_k I_k(s,a) - H_{(s,a),\pi_k}\right| &\le \zeta, \\ \left|\mathbb{E}_k[I_k(s,a)(\tilde{q}_k(s,a) - q^\dagger(s,a))] - F_{(s,a),\pi_k}\right| &\le L\zeta. \end{aligned} \tag{45}$$

Proof. To compare reference and actual episodes, generate the coupled reference and actual episodes together as follows. First generate reference visit indicators $I(s,a)$ and returns $\tilde{q}(s,a)$ from the table. Independently, sample the start component and, for policy π_*, whether the initial bad action survives. If the uniform component is not selected and the initial bad action survives when $\pi = \pi_*$, take the actual visit indicators and first-visit returns to equal these reference values. Otherwise, generate a separate sample for the actual episode from the appropriate conditional distribution. In either case, the reference indicators and returns remain as initially sampled; generating the actual sample does not alter them. Thus the reference sample has the tabulated distribution, and the actual sample has the specified distribution of actual visits and returns. The two samples can disagree only when the uniform component is selected or the initial bad action terminates under $\bar{\sigma}^{(\pi_*)}$. Conditional on the history before the episode, the probability of a disagreement is therefore at most $\epsilon_{\mathrm{ES}} + \tau$.

Each visit indicator $I(s,a)$ takes values zero or one, so each product $I(s,a)(\tilde{q}(s,a) - q^\dagger(s,a))$ lies in $[-L - q^\dagger(s,a), -q^\dagger(s,a)]$, an interval of length L containing zero. Under this coupling, their expectations can therefore differ by at most ζ and $L\zeta$, respectively. ■

D.3 Bounding estimates and counts using the action gaps

Assume that the initial estimates and counts at time k_0 satisfy Theorem 3. For each policy define its cumulative weight

$$\omega_k(\pi) = \frac{k_0 w^\dagger(\pi) + \sum_{t=k_0}^{k-1} \mathbf{1}_{\{\pi_t=\pi\}}}{k}, \qquad \pi \in \mathcal{P}. \tag{46}$$

These weights are nonnegative and sum to one. Since $w^\dagger(\bar{\pi}) = 0$, $\omega_k(\bar{\pi}) = k^{-1} \sum_{t=k_0}^{k-1} \mathbf{1}_{\{\pi_t=\bar{\pi}\}}$. Put $e_k = q_k - q^\dagger$ and define the centered sums

$$\begin{aligned} M_k^I(s,a) &= \sum_{t=k_0}^{k-1} (I_t(s,a) - \mathbb{E}_t I_t(s,a)), \\ M_k^{\tilde{q}}(s,a) &= \sum_{t=k_0}^{k-1} \big(I_t(s,a)(\tilde{q}_t(s,a) - q^\dagger(s,a)) - \mathbb{E}_t[I_t(s,a)(\tilde{q}_t(s,a) - q^\dagger(s,a))]\big). \end{aligned} \tag{47}$$

Each increment has conditional mean zero. We will bound these eight sums simultaneously. Multiplying the sample-average update by its count gives the exact identity

$$n_{k+1}(s,a) q_{k+1}(s,a) = n_k(s,a) q_k(s,a) + I_k(s,a) \tilde{q}_k(s,a).$$

Together with (44), errors ξ_k and ψ_k are implicitly defined by

$$c_k = H\omega_k + \xi_k, \qquad c_k \odot e_k = F\omega_k + \psi_k. \tag{48}$$

For each pair, the errors decompose as

$$\begin{aligned} k\xi_k(s,a) &= n_{k_0}(s,a) - k_0 c^\dagger(s,a) + M_k^I(s,a) \\ &\quad + \sum_{t=k_0}^{k-1} (\mathbb{E}_t I_t(s,a) - H_{(s,a),\pi_t}), \\ k\psi_k(s,a) &= M_k^{\tilde{q}}(s,a) \\ &\quad + \sum_{t=k_0}^{k-1} \big(\mathbb{E}_t[I_t(s,a)(\tilde{q}_t(s,a) - q^\dagger(s,a))] - F_{(s,a),\pi_t}\big). \end{aligned}$$

The initial count error has norm at most one. There is no initial term in the second identity because $q_{k_0} = q^\dagger$. Each bias sum has $k - k_0$ terms, and $(k - k_0)/k \le 1$. Let $u \ge 0$ be a constant. If all eight centered sums in (47) have magnitude at most uk, Lemma 3 gives

$$\|\xi_k\|_\infty \le u + \zeta + 1/k, \qquad \|\psi_k\|_\infty \le u + L\zeta. \tag{49}$$

Equation (63) below bounds the conditional probability that the eight centered-sum bounds fail at any $k \ge k_0$. The two gaps alone do not determine all four estimate errors, since both estimates at a state can change by the same amount. We use the count and return identities to obtain the remaining relations. The matrix D_0 takes the bad-minus-good difference at each state. Define

$$D_0 = \begin{pmatrix} -1 & 1 & 0 & 0 \\ 0 & 0 & -1 & 1 \end{pmatrix}, \qquad d_k = D_0 e_k, \qquad R_k = \max\{\|e_k\|_\infty, \|c_k - c^\dagger\|_\infty\}.$$

Because the reference estimates tie at each state,

$$d_{k,r} = q_k(s_r, \bar{a}) - q_k(s_r, a_*), \qquad r = 1, 2.$$

A positive gap selects the bad action; a nonpositive gap selects the good action, including a tie. The following lemma then provides a deterministic pointwise bound on the difference in the count ratios and value estimates and their reference values in terms of the current action gaps and the cumulative centered sums in (47) and frequency of visiting $\bar{\pi}$. Later lemmas then provide probabilistic bounds on these quantities. The proof involves numerous matrix operations, but the basic idea is simple. Using the reference count and return identities, we first recover the cumulative relative frequency of π_*, π_1, π_2 from the two action gaps and the fact that all policy weights sum to one, keeping the weight of $\bar{\pi}$ and the error terms separate. The bounds in (49) control the departure from these reference identities. Substituting the recovered policy weights back into the identities then bounds the count and estimate errors. Using actual count ratios rather than their reference values introduces a product of the count and estimate errors; this quadratic term is then absorbed into the left-hand side to give the stated bound. We use the larger radius $\ell = 10^{-5}$, twice the target radius, so that the bound can also be applied at a possible first exit.

Lemma 4 (Pointwise bound)**.** *Put* $\kappa = (1 - (304/3)\ell)^{-1} = 18750/18731$*. If* $R_k \le \ell$ *and* (49) *holds, then*

$$R_k \le \kappa\big(10\,\|d_k\|_\infty + 35\omega_k(\bar\pi) + (32L+1)\zeta + 33u + 1/k\big). \tag{50}$$

Proof. We first solve for the weights of policies π_*, π_1, π_2 using their total weight and the two action gaps, keeping the weight of $\bar\pi$ and the errors as separate inputs. Substitution into the count and return-sum identities then recovers the estimate and count errors. Let I_4 be the 4×4 identity matrix and $\mathbf{1}_3 = (1,1,1)^\top$. Set $C_\dagger = \operatorname{diag}(c^\dagger)$, $U = C_\dagger^{-1}F$. Multiplication by $C_\dagger^{-1}$ divides each sum of return differences by its reference visit frequency, so the columns of U give the corresponding contributions after dividing by the reference visit frequencies. Write $U_{\mathrm{r}}, H_{\mathrm{r}}$ for their columns labelled π_*, π_1, π_2, in that order; $U_{\cdot,\bar\pi}, H_{\cdot,\bar\pi}$ are the remaining columns. The matrix

$$M = \begin{pmatrix} \mathbf{1}_3^\top \\ D_0 U_{\mathrm{r}} \end{pmatrix} \tag{51}$$

maps these three weights to their total weight and two gap contributions. Its determinant is $-1155904/3132675 \neq 0$, so these three quantities determine the weights. Set

$$T_{\mathrm{sel}} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad m_{\bar\pi} = \begin{pmatrix} 1 \\ D_0 U_{\cdot,\bar\pi} \end{pmatrix}.$$

The matrix T_{sel} inserts the two gap coordinates below a zero total-weight coordinate; the vector $m_{\bar\pi}$ records policy $\bar\pi$'s contribution to total weight and gaps. Define

$$A = U_{\mathrm{r}} M^{-1} T_{\mathrm{sel}}, \quad B = H_{\mathrm{r}} M^{-1} T_{\mathrm{sel}}.$$

Thus A and B give the gap contributions to the recovered estimate and normalized count errors, respectively. Put

$$b = U_{\cdot,\bar\pi} - U_{\mathrm{r}} M^{-1} m_{\bar\pi}, \qquad v = H_{\cdot,\bar\pi} - H_{\mathrm{r}} M^{-1} m_{\bar\pi}.$$

These vectors give the net effects of policy $\bar\pi$ on estimates and counts after the other three weights compensate for its total-weight and gap contributions. To account for using reference counts in $C_\dagger$ rather than actual counts, set $\chi_k = \operatorname{diag}((c_k - c^\dagger)/c^\dagger)e_k$, where division is coordinatewise. Then $C_\dagger^{-1}(c_k \odot e_k) = e_k + \chi_k$ and $\|\chi_k\|_\infty \le (16/3)R_k^2$. Each coordinate is a product of a count error and an estimate error, which is why this correction is quadratic. Apply $D_0 C_\dagger^{-1}$ to the second identity in (48). Together with $\sum_{\pi\in\mathcal{P}} \omega_k(\pi) = 1$, this gives a linear system with matrix M for the weights labelled π_*, π_1, π_2. Subtracting the corresponding reference weights gives, with $\omega_{k,\mathrm{r}} = (\omega_k(\pi_*), \omega_k(\pi_1), \omega_k(\pi_2))^\top$ and $w_{\mathrm{r}}^\dagger = (w^\dagger(\pi_*), w^\dagger(\pi_1), w^\dagger(\pi_2))^\top$,

$$M(\omega_{k,\mathrm{r}} - w_{\mathrm{r}}^\dagger) = T_{\mathrm{sel}}(d_k + D_0\chi_k - D_0 C_\dagger^{-1}\psi_k) - m_{\bar\pi}\omega_k(\bar\pi).$$

The first row uses $\sum_{\pi\in\mathcal{P}} \omega_k(\pi) = 1$ and the other two use $d_k = D_0 e_k$. Inverting M and substituting into (48) gives

$$\begin{aligned} e_k &= Ad_k + (AD_0 - I_4)\chi_k + b\omega_k(\bar\pi) + (C_\dagger^{-1} - AD_0C_\dagger^{-1})\psi_k, \\ c_k - c^\dagger &= Bd_k + BD_0\chi_k + v\omega_k(\bar\pi) - BD_0C_\dagger^{-1}\psi_k + \xi_k. \end{aligned} \tag{52}$$

Here $AD_0 - I_4$ and BD_0 carry the quadratic correction into the estimate and count errors. The coefficients of ψ_k carry the return error through the recovered weights and include its direct contribution $C_\dagger^{-1}\psi_k$ to the estimates. The count error ξ_k enters directly. Direct calculation gives the following norms to three significant figures:

$$\begin{aligned} \|A\|_\infty &\approx 9.62, & \|B\|_\infty &\approx 5.49, \\ \|AD_0 - I_4\|_\infty &\approx 18.2, & \|BD_0\|_\infty &\approx 11.0, \\ \|b\|_\infty &\approx 35.0, & \|v\|_\infty &\approx 21.1, \\ \left\|C_\dagger^{-1} - AD_0C_\dagger^{-1}\right\|_\infty &\approx 31.2, & \left\|BD_0C_\dagger^{-1}\right\|_\infty &\approx 20.2. \end{aligned}$$

Taking maxima in pairs gives bounds of 10 for gap terms, 19 for the quadratic product, 35 for the weight of $\bar\pi$, and 32 for the return error. Thus

$$R_k \le 10\,\|d_k\|_\infty + 35\omega_k(\bar\pi) + (32L+1)\zeta + 33u + 1/k + (304/3)R_k^2.$$

The return and count errors give $(32L+1)\zeta$ and $33u = 32u + u$. Since $R_k \le \ell$, the last term is at most $(304/3)\ell R_k$. Rearranging gives the stated factor κ. ∎

D.4 Bounding how often the all-bad policy is used

The all-bad policy $\bar{\pi}$ has zero weight in the reference process. We bound the number of episodes using this policy in terms of the number with a uniform start, to control the $\omega_k(\bar{\pi})$ term in Lemma 4. Write $Z_k = kd_k = (X_k, Y_k)$. For a single update, abbreviate $(X, Y) = Z_k$ and $(X', Y') = Z_{k+1}$, and put $\Delta X = X' - X$, $\Delta Y = Y' - Y$. Assume throughout this subsection that each $q_k(s,a)$ and $\widehat{c}_k(s,a)$ is within ℓ of its reference value. Then $\|d_k\|_\infty \le 2\ell$. Suppressing k on the right,

$$\begin{aligned}\Delta X &= d_{k,1} + \frac{I(s_1,\bar{a})(\tilde{q}(s_1,\bar{a}) - q(s_1,\bar{a}))}{\widehat{c}(s_1,\bar{a})} - \frac{I(s_1,a_*)(\tilde{q}(s_1,a_*) - q(s_1,a_*))}{\widehat{c}(s_1,a_*)},\\ \Delta Y &= d_{k,2} + \frac{I(s_2,\bar{a})(\tilde{q}(s_2,\bar{a}) - q(s_2,\bar{a}))}{\widehat{c}(s_2,\bar{a})} - \frac{I(s_2,a_*)(\tilde{q}(s_2,a_*) - q(s_2,a_*))}{\widehat{c}(s_2,a_*)}.\end{aligned} \tag{53}$$

The term d_k appears because the scaling factor changes from k to $k+1$. Substitution of the stated ranges into (53) gives the following coordinatewise upper bounds, rounded upwards to three significant figures:

Policy	Class	Upper increment (3 s.f.)
π_*	A or B	$(0.723, 0)$
π_*	C	$(-0.166, 0.0555)$
π_1	F	$(-0.166, -0.239)$
π_1	G	$(3.55, -0.973)$
π_2	D	$(-0.166, 0)$
π_2	E	$(-3.61, 0.961)$
$\bar{\pi}$	Start at $(s_2, \bar{a})$	$(0.0000200, -5.19)$

The calculation uses the sign of each gap under the indicated policy. For example, in A or B the policy is π_*, so $d_{k,1} \le 0$ and

$$\Delta X \le \frac{1/6 + \ell}{3/16 - \ell} - \frac{1/6 - \ell}{1 + \ell} < 0.723.$$

Class A leaves the other gap unchanged, and B decreases it, giving the common upper bound 0 in that column. The other rows follow by the same substitution into (53).

These bounds hold throughout the stated ranges and include the d_k term in (53). Under policy π_*, if the initial bad action results in termination then its estimate decreases, leaving both gaps non-positive. These episodes are covered separately because the reference law conditions on that action on not terminating.

We also need an upper bound that holds for every initial pair. Fix a state s and suppress the episode index. Under the bounds stated above, the estimates are negative and $\widehat{c}(s,\bar{a}) < \widehat{c}(s,a_*)$. If an episode visits both pairs at that state, then $\tilde{q}(s,a_*) = \tilde{q}(s,\bar{a}) \le 0$, and their contribution in (53) is

$$\tilde{q}(s,a_*)\left(\frac{1}{\widehat{c}(s,\bar{a})} - \frac{1}{\widehat{c}(s,a_*)}\right) - \frac{q(s,\bar{a})}{\widehat{c}(s,\bar{a})} + \frac{q(s,a_*)}{\widehat{c}(s,a_*)} \le -\frac{q(s,\bar{a})}{\widehat{c}(s,\bar{a})}.$$

The same upper bound holds if the episode visits the bad-action pair but not the good-action pair at that state. If neither pair is visited, the contribution is zero. If the episode visits (s_1,a_*) but not $(s_1,\bar{a})$, use $\tilde{q}(s_1,a_*) \ge -L$. If it visits (s_2,a_*) but not $(s_2,\bar{a})$, the stronger bound $\tilde{q}(s_2,a_*) \ge -1$ holds, since a terminal loss of $-L$ would require visiting $(s_2,\bar{a})$. Adding the term from the change in scaling gives

$$\begin{aligned}\Delta X &\le 2\ell + \max\left\{\frac{1/6+\ell}{3/16-\ell}, \frac{L - 1/6 + \ell}{1-\ell}\right\} < 3.547685 < 3.55,\\ \Delta Y &\le 2\ell + \max\left\{\frac{1/5+\ell}{65/96-\ell}, \frac{4/5+\ell}{5/6-\ell}\right\} < 0.960044 < 0.961.\end{aligned}$$

We use a nonnegative function that, on an episode with a nonuniform start, decreases by at least one under policy $\bar{\pi}$ and cannot increase under the other policies. Its increase on a uniform-start episode is bounded. Summing its changes will give the required bound.

Let $(z)_+ = \max\{z, 0\}$, put $a = 71/20 = 3.55$, $b_0 = 961/1000 = 0.961$, and define

$$\Psi(X, Y) = 20(X - a)_+ + 100(Y - b_0)_+ + \frac{11}{10}\mathbf{1}_{\{X>0, Y>0\}}. \tag{54}$$

Write $\Delta\Psi = \Psi(X', Y') - \Psi(X, Y)$ for its change in one update. The function is nonnegative, nondecreasing in each coordinate, and zero at the origin. Its indicator term is present precisely when policy $\bar{\pi}$ is selected.

Lemma 5 (A bound on the use of policy $\bar{\pi}$)**.** *Let $U_k = 1$ if episode k selects the uniform-start component, and $U_k = 0$ if it selects $\bar{\sigma}^{(\pi_k)}$ instead. Whenever the bounds on $q_k, \widehat{c}_k$ stated above hold before episode k,*

$$\mathbf{1}_{\{\pi_k = \bar{\pi}\}} + \Psi(Z_{k+1}) - \Psi(Z_k) \le (846/5)U_k = 169.2U_k. \tag{55}$$

If $d_{k_0} = 0$ and these bounds hold before all episodes through $k - 1$, then

$$\omega_k(\bar{\pi}) \le \frac{846}{5k} \sum_{t=k_0}^{k-1} U_t. \tag{56}$$

Proof. Since Ψ is nondecreasing in both coordinates, replacing the actual increments by their upper bounds in the table gives an upper bound on $\Psi(Z_{k+1})$. First take $U_k = 0$.

Policy π_* has $X, Y \le 0$, so all three terms of Ψ are zero. Classes A/B and C keep these terms zero, as do episodes in which the initial bad action terminates immediately. Policy π_2 has $X > 0, Y \le 0$. Class D cannot increase Ψ. In E, $X' \le X - 3.61$ and $Y' \le Y + b_0 \le b_0$. If it enters the positive quadrant, then $X > 3.61$ and the horizontal linear term decreases by at least $20(3.61 - a) = 1.2$. This exceeds the increase of 1.1 in the indicator term. Otherwise no term increases.

Policy π_1 has $X \le 0, Y > 0$. Class F cannot increase Ψ. In G, $X' \le a$ and $Y' \le Y - 0.973$. Entry into the positive quadrant requires $Y > 0.973$, so the vertical term decreases by at least $100(0.973 - b_0) = 1.2$. Otherwise no term increases.

Policy $\bar{\pi}$ has $X, Y > 0$. Starting at $(s_2, \bar{a})$ gives $X' \le X + 0.0000200$ and $Y' \le Y - 5.19$, directly from the table. The horizontal term increases by at most $20(0.0000200) = 0.0004$. Leaving the quadrant removes the indicator, so $\Delta\Psi \le 0.0004 - 1.1 = -1.0996 \le -1$. Remaining inside requires $Y > 5.19$, so the vertical term decreases by at least $100(5.19 - b_0) = 422.9$. Then $\Delta\Psi \le 0.0004 - 422.9 = -422.8996 < -1$.

For $U_k = 1$, use the bounds that hold for every initial pair:

$$\mathbf{1}_{\{\pi_k = \bar{\pi}\}} + \Delta\Psi \le 1 + 20a + 100b_0 + 11/10 = 1 + 71 + 96.1 + 1.1 = 169.2.$$

This proves (55) for every actual episode. Summing from k_0 to $k - 1$, using $\Psi(Z_{k_0}) = 0$ and $\Psi \ge 0$, and dividing by k gives (56). ■

D.5 Probability bounds for all subsequent episodes

We obtain bounds that hold simultaneously at every later step, so that they can be applied at a possible first exit from the trapping region.

An exponential bound. Let V_t be $\mathcal{F}_{t+1}$-measurable, with $\mathbb{E}_t V_t = 0$ and $|V_t| \le b_t$, where $b_t \ge 0$ is deterministic. For fixed $\lambda \in \mathbb{R}$, the process

$$E_k(\lambda) = \exp\left\{\lambda \sum_{t=k_0}^{k-1} V_t - \frac{\lambda^2}{2} \sum_{t=k_0}^{k-1} b_t^2\right\} \tag{57}$$

starts at one and satisfies $\mathbb{E}_k E_{k+1}(\lambda) \le E_k(\lambda)$, so it is a nonnegative supermartingale.

To prove the supermartingale property, first suppose $b_t > 0$. Convexity gives

$$e^{\lambda v} \le \frac{b_t + v}{2b_t} e^{\lambda b_t} + \frac{b_t - v}{2b_t} e^{-\lambda b_t}, \qquad -b_t \le v \le b_t.$$

Conditional expectation and $\mathbb{E}_t V_t = 0$ give

$$\mathbb{E}_t e^{\lambda V_t} \le \frac{e^{\lambda b_t} + e^{-\lambda b_t}}{2} = \cosh(\lambda b_t) \le e^{\lambda^2 b_t^2/2}.$$

The last inequality follows by comparing coefficients, using $(2m)! \ge 2^m m!$. For $b_t = 0$, $V_t = 0$ and the same bound holds. Multiplication by $e^{-\lambda^2 b_t^2/2}$ gives the required conditional inequality.

For any nonnegative supermartingale (S_k) starting with $S_{k_0} = 1$ and any threshold $A > 0$,

$$\mathbb{P}(\exists k \ge k_0 : S_k \ge A \mid \mathcal{F}_{k_0}) \le 1/A. \tag{58}$$

To prove this, fix $N \ge k_0$ and stop the process at its first crossing of A, or at N if there is no crossing. Iterating the conditional inequality gives an expected stopped value at most one, conditional on $\mathcal{F}_{k_0}$. This value is at least A on crossing paths and nonnegative otherwise, so $A\,\mathbb{P}(\text{crossing by } N \mid \mathcal{F}_{k_0}) \le 1$. Letting $N \to \infty$ gives (58). This is the nonnegative-supermartingale maximal inequality; see [15] for broader context.

Bounds on one update. Write $\Delta q_k = q_{k+1} - q_k$, and similarly for c, d. Initially $n_{k_0}(s,a) \le k_0 c^\dagger(s,a) + 1 \le k_0 + 1$, and each later episode adds at most one count. Thus $0 \le n_k(s,a) \le k+1$, so $0 \le c_k(s,a) \le 1 + 1/k$. The count update therefore gives

$$\|(k+1)\Delta c_k\|_\infty = \|I_k - c_k\|_\infty \le \max_{s,a} \max\{1, c_k(s,a)\} \le 1 + 1/k < 2. \tag{59}$$

Also $\widehat{c}_k(s,a) - c_k(s,a) = (1 - c_k(s,a))/(k+1)$, whose magnitude is at most $1/(k+1)$.

For the estimate and gap bounds, suppose $q_k, \widehat{c}_k$ are within ℓ of their reference values. Since $q^\dagger(s,a)$ is either $-1/6$ or $-1/5$ and $\ell = 10^{-5}$, we have $-1 < q^\dagger(s,a) - \ell \le q_k(s,a) \le q^\dagger(s,a) + \ell < 0$. The scaled estimate update is

$$(k+1)\Delta q_k(s,a) = \frac{I_k(s,a)(\tilde{q}_k(s,a) - q_k(s,a))}{\widehat{c}_k(s,a)}, \qquad \widehat{c}_k(s,a) = \frac{n_k(s,a)+1}{k+1}.$$

Its denominator satisfies $\widehat{c}_k(s,a) \ge c^\dagger(s,a) - \ell \ge 3/16 - \ell > 1/6$. Put

$$B(s,a) = \frac{L + q^\dagger(s,a) + \ell}{c^\dagger(s,a) - \ell}.$$

Since $\tilde{q}_k(s,a) \in \{0, -1, -L\}$, this scaled update has magnitude at most $B(s,a)$. Substitution gives the following strict upper bounds, rounded upwards:

(s,a)	(s_1,a_*)	$(s_1,\bar{a})$	(s_2,a_*)	$(s_2,\bar{a})$
$B(s,a) <$	3.548	18.922	4.218	5.191

The two scaled estimate updates at a state have the same sign. An unvisited pair contributes zero. For visited pairs, the returns are equal: a return of zero gives a nonnegative update, while a return of -1 or $-L$ gives a nonpositive update. Their difference is therefore bounded by their larger magnitude, rather than their sum. Consequently

$$\|(k+1)\Delta q_k\|_\infty < 19, \qquad |(k+1)\Delta d_{k,1}| < 19, \qquad |(k+1)\Delta d_{k,2}| < 26/5. \tag{60}$$

At a fixed state s, write $B_{a_*} = B(s,a_*)$ and $B_{\bar{a}} = B(s,\bar{a})$. Every scaled gap increment belongs to $[-B_{\bar{a}}, B_{a_*}]$. For an episode with a negative terminal reward, the two scaled estimate updates belong to $[-B_{a_*}, 0]$ and $[-B_{\bar{a}}, 0]$. For terminal reward zero, each scaled update is at most $(-q^\dagger(s,a) + \ell)/(c^\dagger(s,a) - \ell) < 1$, whereas every $B(s,a) > 3.5$, so the same interval contains their difference. The conditional mean also belongs to this interval. Thus

$$\begin{aligned} |(k+1)(\Delta d_{k,1} - \mathbb{E}_k \Delta d_{k,1})| &\le B(s_1,a_*) + B(s_1,\bar{a}) < 22.470 < 23, \\ |(k+1)(\Delta d_{k,2} - \mathbb{E}_k \Delta d_{k,2})| &\le B(s_2,a_*) + B(s_2,\bar{a}) < 9.408 < 10. \end{aligned} \tag{61}$$

These bound the centered increments; (60) bounds the increments before centering.

Signs of the expected gap changes. For fixed $q_k, \widehat{c}_k$, the reference mean scaled estimate increment is

$$\frac{\mathbb{E}_\pi^{\mathrm{ref}}[I(s,a)\tilde{q}(s,a)] - H_{(s,a),\pi} q_k(s,a)}{\widehat{c}_k(s,a)}.$$

Subtract its good coordinate from its bad coordinate at each state. Bounding the resulting expressions over the stated ranges of q_k and $\widehat{c}_k$ gives:

π	reference mean of $(k+1)\Delta d_{k,1}$	reference mean of $(k+1)\Delta d_{k,2}$
π_*	$[.007593, .007657]$	$[.028810, .028867]$
π_1	$[1.071405, 1.071452]$	$[-.484457, -.484334]$
π_2	$[-1.314926, -1.314704]$	$[.319992, .320008]$

The endpoints are rounded outwards. A negative gap selects the good action and has a positive expected change; a positive gap selects the bad action and has a negative expected change.

The coupling in Lemma 3 also applies to these increments with the current $q, \widehat{c}$ fixed. On a disagreement, their difference is bounded by the interval widths in (61). The actual and reference means therefore differ by at most 23ζ and 10ζ. Under (36), $23\zeta \le 2.3092 \times 10^{-7}$, smaller than .007593, the smallest magnitude in the table. The signs are therefore preserved for actual episodes.

For policy $\bar{\pi}$, both gaps are positive and a direct conditioning on the four initial pairs gives

$$\mathbb{E}_k[(k+1)\Delta d_{k,1}] = \frac{\epsilon_{\mathrm{ES}}}{4}\left\{\frac{2(-1-q_k(s_1,\bar{a}))}{\widehat{c}_k(s_1,\bar{a})} - \frac{-L/3 - q_k(s_1,a_*)}{\widehat{c}_k(s_1,a_*)}\right\},$$

$$\mathbb{E}_k[(k+1)\Delta d_{k,2}] = \frac{(1-2\epsilon_{\mathrm{ES}}/3)(-L-q_k(s_2,\bar{a}))}{\widehat{c}_k(s_2,\bar{a})} - \frac{\epsilon_{\mathrm{ES}}(-1-q_k(s_2,a_*))}{4\widehat{c}_k(s_2,a_*)}.$$

Over the same ranges, the first bracket is below -7.81 and the second mean is below -5.19. Since $\epsilon_{\mathrm{ES}} > 0$, both means are negative. Thus, whenever the stated coordinate bounds hold,

$$d_{k,r} > 0 \Longrightarrow \mathbb{E}_k \Delta d_{k,r} \le 0, \qquad d_{k,r} < 0 \Longrightarrow \mathbb{E}_k \Delta d_{k,r} \ge 0. \tag{62}$$

No drift sign is required when a gap is zero.

The first possible exit. Use u, δ from Theorem 3, set $R_0 = 5 \times 10^{-6}$, and define

$$k_{\mathrm{exit}} = \inf\{k \ge k_0 : R_k \ge R_0\},$$

with $\inf \varnothing = \infty$. Before this step, $\left\|\widehat{c}_k - c^\dagger\right\|_\infty \le R_0 + 1/(k+1) < \ell$, so all local bounds above apply. The stopped gap sums below use $\mathbf{1}_{\{t<k_{\mathrm{exit}}\}}$, which is $\mathcal{F}_t$-measurable. Their increments therefore retain conditional mean zero and include the update from $k_{\mathrm{exit}} - 1$ to k_{exit}, which would first violate $R_k < R_0$.

The eight count and return sums. Since $I_t(s,a) \in \{0,1\}$ and $0 \le \mathbb{E}_t I_t(s,a) \le 1$, $|I_t(s,a) - \mathbb{E}_t I_t(s,a)| \le 1$. A centered return increment has magnitude at most L, because $I(s,a)(\tilde{q}(s,a) - q^\dagger(s,a))$ lies in an interval of length L. For either kind, let b be its bound, and apply (57) with $\lambda = 2u/b^2$. If its sum M_k exceeds uk, then

$$\log E_k(\lambda) > \lambda u k - \frac{\lambda^2 b^2}{2}(k - k_0) = \frac{2k_0 u^2}{b^2}.$$

The maximal inequality gives a crossing probability at most $e^{-2k_0u^2/b^2}$. Apply it to both signs of all four coordinates. Adding the failure probabilities, with $b = 1$ for counts and $b = L$ for returns, gives

$$\begin{aligned}\mathbb{P}(\exists k \ge k_0, s, a :\ & |M_k^I(s,a)| > uk \text{ or } |M_k^{\tilde{q}}(s,a)| > uk \mid \mathcal{F}_{k_0}) \\ & \le 8e^{-2k_0u^2} + 8e^{-2k_0u^2/L^2}.\end{aligned} \tag{63}$$

This bound holds without stopping or local assumptions. Outside this failure event, (49) therefore holds for every $k \ge k_0$, including a possible first exit.

Bounds on the gaps. For $r = 1, 2$ define the stopped centered sum

$$M^d_{k,r} = \sum_{t=k_0}^{k-1} \mathbf{1}_{\{t<k_{\text{exit}}\}} \big(\Delta d_{t,r} - \mathbb{E}_t \Delta d_{t,r}\big).$$

By (61), its increment bounds are $23/(t+1)$ and $10/(t+1)$. Their squared sums over the whole future are at most $23^2/k_0$ and $10^2/k_0$, since

$$\sum_{t=k_0}^{\infty} (t+1)^{-2} \le \int_{k_0}^{\infty} x^{-2}\, dx = 1/k_0.$$

For a sum of squared bounds at most $B > 0$, a crossing $M^d_{k,r} > \delta$ forces $\log E_k(\lambda) > \lambda\delta - \lambda^2 B/2$. Choosing $\lambda = \delta/B$ and applying the maximal inequality gives probability at most $e^{-\delta^2/(2B)}$. Include the negative crossing as well. Thus both stopped sums remain within δ at every later step except on an event of conditional probability at most

$$2e^{-k_0\delta^2/1058} + 2e^{-k_0\delta^2/200}, \qquad 1058 = 2 \cdot 23^2, \quad 200 = 2 \cdot 10^2. \tag{64}$$

On the event that both stopped gap sums stay within δ at every later step, a sum over any interval has magnitude at most 2δ, as the difference of two partial sums. Suppose $k_0 < k \le k_{\text{exit}}$ and $d_{k,r} > 0$. Let m be the last index in $[k_0, k-1]$ with $d_{m,r} \le 0$; it exists because $d_{k_0} = 0$. The crossing step gives $d_{m+1,r} \le 19/(m+1) \le 19/(k_0+1)$, using the larger of the two gap bounds in (60). All subsequent gaps before episodes $m+1, \dots, k-1$ are positive, so their conditional mean increments are nonpositive by (62). Consequently

$$\begin{aligned} d_{k,r} &= d_{m+1,r} + \sum_{t=m+1}^{k-1} \mathbb{E}_t \Delta d_{t,r} + M^d_{k,r} - M^d_{m+1,r} \\ &\le \frac{19}{k_0+1} + 2\delta. \end{aligned}$$

The last nonnegative index gives the corresponding lower bound when $d_{k,r} < 0$. The choice of m may depend on the observed path because all partial-sum bounds hold on the same event. Hence

$$\sup_{k_0 \le k \le k_{\text{exit}}} \|d_k\|_\infty \le 19/(k_0+1) + 2\delta. \tag{65}$$

The number of uniform starts. Put $\bar{p} = 4 \times 10^{-11}$. The count $C^U_k = \sum_{t=k_0}^{k-1} U_t$ is the number of uniform starts. Since $U_t \in \{0, 1\}$ and $\mathbb{P}(U_t = 1 \mid \mathcal{F}_t) = \epsilon_{\text{ES}} \le \bar{p}$, we have $\mathbb{E}_t[2^{U_t}] = 1 + \epsilon_{\text{ES}} \le e^{\bar{p}}$, so

$$\exp\{(\log 2) C^U_k - \bar{p}(k - k_0)\}$$

is another nonnegative supermartingale starting at one. If $C^U_k > 2\bar{p}k$, its logarithm exceeds

$$\bar{p}\{(2\log 2 - 1)k + k_0\} \ge 2(\log 2)\bar{p}k_0,$$

using $k \ge k_0$ and $2\log 2 - 1 > 0$. Therefore

$$\mathbb{P}(\exists k \ge k_0 : C^U_k > 2\bar{p}k \mid \mathcal{F}_{k_0}) \le e^{-2(\log 2)\bar{p}k_0}. \tag{66}$$

Let $\mathcal{A}$ be the event that the eight sums $M^I_k(s,a), M^{\tilde{q}}_k(s,a)$ satisfy the uk bounds, both stopped gap sums satisfy the δ bounds, and $C^U_k \le 2\bar{p}k$, for every $k \ge k_0$. Adding (63), (64) and (66) gives

$$\mathbb{P}(\mathcal{A} \mid \mathcal{F}_{k_0}) \ge 1 - \beta_{\text{trap}}(k_0). \tag{67}$$

These bounds hold simultaneously at every later step, so they apply at the possibly random first exit used below.

D.6 Trapping

Proof of Theorem 3. On $\mathcal{A}$ there can be no first exit. Initially $R_{k_0} \le 1/k_0 < R_0$ and $d_{k_0} = 0$. Suppose the first exit k_{exit} is finite. At the preceding step, $R_{k_{\mathrm{exit}}-1} < R_0$. Equations (59)–(60) bound the update to k_{exit}, giving

$$R_{k_{\mathrm{exit}}} \le R_0 + 19/(k_0+1) < 10^{-5}.$$

Thus Lemma 4 can be applied at k_{exit}.

Since $d_{k_0} = 0$, we have $\Psi(Z_{k_0}) = 0$, and all preceding episodes satisfy the local bounds in Lemma 5. Together with the uniform-start bound this gives

$$\omega_{k_{\mathrm{exit}}}(\bar{\pi}) \le \frac{846}{5}(2\bar{p}) = 338.4\bar{p}.$$

The count and return identities and the stopped noise bounds include the update to k_{exit}. Lemma 4 and (65) therefore give

$$R_{k_{\mathrm{exit}}} \le \kappa\Bigg[10\left(\frac{19}{k_0+1} + 2\delta\right) + 35\left(\frac{846}{5}\,2\bar{p}\right) + (32L+1)(10^{-8}+\bar{p}) + 33u + \frac{1}{k_0}\Bigg] < 4.943 \times 10^{-6} < R_0.$$

The numerical inequality uses $k_0 \ge 10^{17}$, $\bar{p} = 4 \times 10^{-11}$, $u = 2 \times 10^{-8}$, $\delta = 1.3 \times 10^{-7}$ and $\kappa = 18750/18731$; each k_0-dependent term decreases with k_0. It contradicts $R_{k_{\mathrm{exit}}} \ge R_0$. Hence $R_k < R_0$ for every $k \ge k_0$, and (67) gives its conditional probability.

The optimal good-action value at s_2 is zero, whereas its reference estimate is $-1/5$. The tube therefore keeps the estimates at distance greater than $1/5 - R_0$ from the optimal vector. ■

Reaching the prescribed estimates and counts. We now give two sequences consisting of episodes taken from those labelled A–G in Appendix D.2. Following the initial sequence, and then repetitions of the second sequence, reaches the starting point for Theorem 3 at arbitrarily late episode indices. Each sequence has integer visit counts in the same ratio as $c^\dagger$ and, for each pair, a ratio of total returns to counts equal to $q^\dagger$. The initial sequence starts from empty return lists with $q_0 = (-L, 0, -L, -L)$; the second starts at $q^\dagger$ with counts contributed by one or more completed sequences.

The count and return constraints already require a minimum length of 192 episodes for such a sequence. To see this, every A–G episode visits (s_1, a_*), so a sequence of r episodes with the prescribed count ratios has $65r/96$ visits to $(s_2, \bar{a})$. Only class G contributes a nonzero return at this pair, so its required average $-1/5$ forces exactly $(65r/96)/(5L) = 7r/192$ class G episodes. This is an integer, so 192 divides r. In addition, the greedy policy before each episode must correspond to the indicated class. A length of 192 episodes is also sufficient, and there are, in fact, many possible such sequences. We give one pair below. Each adds return sums $-(32, 6, 32, 26)$ and visit counts $(192, 36, 160, 130)$ in the stated pair order. The initial sequence is

$$\begin{aligned} P = \; & ECFGDC^6FCAC^3GADEC^2FC^4FC^3BCAC^3ACGD^2EC^2FC^4GC^3AC^2B \\ & DC^4ADC^4BEFCFC^4GC^2AC^3AC^2BDC^4AC^3ADEF^2GC^9ACBC^4AC^2FC \\ & BCAD^2C^7FAC^3BC^2AEFGC^7AC^2FCBC^2AC^3ADC^3FC^2BC^4BC^2ADC^4A \end{aligned}$$

and the repeating sequence is

$$W = CGD^{20}EG(C^9A^3)(EG)^5D^2\left(C^5B(C^4B)^2\right)^8(C^9A^3)D^4.$$

Powers denote repetitions. The greedy choices during P are checked directly, and the checks for W are linear in j, so their slopes and values at $j = 1$ establish compatibility for every repetition. These calculations are omitted. There are many alternative sequences with the same effect.

For every integer $m \ge 1$, the finite history PW^{m-1} therefore reaches $q_{192m} = q^\dagger$ and $n_{192m} = 192mc^\dagger$, so $c_{192m} = c^\dagger$. Each such history has positive probability, since every required start and transition has positive probability. We then choose a finite m with $192m \ge 10^{17}$ and $\beta_{\mathrm{trap}}(192m) < 1$, possible as $\beta_{\mathrm{trap}}(k) \to 0$, then condition on this history and apply Theorem 3. Its positive conditional success probability, together with the positive probability of the history, proves Theorem 2.

ChatGPT 6.0 and 6.1 were used to help construct this example and the proof.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Both the abstract and the introduction are an accurate reflection of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The 13-state example is strongly evidenced by simulations, but only proved in noiseless setting. This is made clear. The assumptions of Theorem 1 are also discussed.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: The main paper contains a sketch of the proof of Theorem 1, clearly stating that it is an adaptation of the main convergence result in [13]. Appendix B then describes how the setting of this paper affects that proof. In general, each theorem starts by listing its assumptions.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper includes simulations. Full details of how these were performed are given in Appendices C and D, with sufficient detail to allow them to be reproduced, including the results of intermediate calculations.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The Matlab code used to produce the simulations is provided as supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: The paper does not give experimental results in this sense. Full details of the numerical simulations are provided, as detailed in other answers.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: The paper does not present experimental results in this sense.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: The paper does not present experimental results in this sense. The Matlab simulations were run on a standard laptop.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have read the Code of Ethics, and this work complies in all respects.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational work, far removed from application at this stage.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    - The answer [N/A] means that
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: An LLM was also used to help refine and code the 13 state First-Visit example and to help construct the two state First-Visit example and its proof. This is noted at the relevant points in the Appendix.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.