INSTANCE-DEPENDENT AUTO-EXPLORATION FOR TABULAR ONLINE DISCOUNTED REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In online discounted reinforcement learning from a single trajectory, a central cost is exploration: the time spent reaching states whose optimal action is still unknown. Existing auto-exploration guarantees for stochastic policy mirror descent charge every iteration the cost of reaching the hardest state, even once its optimal action is known, through a constant exponential in the mixing time of the optimal policy. We prove that the exploration cost is controlled instead by a *certified hitting cost* $H_\theta(s)$, which is equivalent up to constants to the worst-case expected time to reach s under policies that keep mass at least θ on optimal actions. For the algorithm's θ, every iterate is such a policy with high probability. Each rollout stops once it covers the states where the policy keeps mass at least θ on two or more actions. Resolved states leave this set, so every iteration is charged only the largest $H_\theta(s)$ over unresolved states. For accuracy ε and advantage gaps $\Delta(s)$, the sample complexity is then at most an ε^{-2} baseline plus the exploration cost $\max_s H_\theta(s)/\max\{(1-\gamma)\varepsilon, \Delta(s)\}^2$, up to logarithmic and instance-independent factors. The only structural assumption is that some deterministic optimal policy induces an irreducible chain; the mixing regime and action-independent transitions are special cases with explicit bounds on H_θ. On an action-dependent family of $3n-2$ states whose small gaps arise from transitions, the upper bound $\widetilde{O}(n^2+n\varepsilon^{-2})$ matches a lower bound with deterministic costs up to logarithmic factors, for a fixed discount factor. Charging the hardest state at every iteration gives $\Omega(n^2\varepsilon^{-2})$.

1 INTRODUCTION

In online discounted reinforcement learning with discount factor γ and finite spaces $\mathcal{S}, \mathcal{A}$, an agent learns from a single trajectory of an unknown MDP, with no resets and no simulator, and must output a policy whose value is within ε of optimal. The auto-exploration framework of Ju & Lan (2025) shows that stochastic policy mirror descent (SPMD) with Tsallis regularization needs $\widetilde{O}(\varepsilon^{-2})$ samples without prior knowledge of mixing parameters. Its exploration cost is governed by one global quantity D_{expl}: every iteration pays the time needed to reach the hardest-to-reach state, even once its optimal action is known, and D_{expl} is exponential in the mixing time of the optimal policy.

This charge conflates how hard a state is to *reach* with how long it takes to *resolve*, that is, to identify its optimal action, which takes about $\Delta(s)^{-2}$ visits for advantage gap $\Delta(s)$. Consider 100 states: a bottleneck visited once every 10^4 steps with gap 0.6, and 99 generic states visited every 100 steps with gap 0.01. At accuracy 0.01 there are about $\varepsilon^{-2} = 10^4$ iterations, and the bottleneck is resolved as soon as the accuracy falls below its gap. Yet the global analysis charges the bottleneck's visit cost at every iteration, a total of order 10^8, whereas charging it only while it is unresolved leaves the cost of the generic states, of order 10^6.

Charging only unresolved states needs a single change to SPMD: a rollout stops once it covers the *active set*, the states with at least two *non-rare* actions (probability at least a threshold $\underline{\pi}$), rather than every state. The active set is a computable stand-in for the unresolved states, so every iteration pays only for reaching the hardest of them.

What has to be closed. Charging each state only while it is unresolved is not a bookkeeping change: three obstacles block the upper bound, and a fourth blocks a matching lower bound. *(i) Hitting under*

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

a moving policy. The rollout length at iteration t depends on the random current policy π_t. Controlling it through the mixing time and stationary distribution of π_t, as in Ju & Lan (2025), costs a factor $\underline{\pi}^{-\Theta(\bar{b})}$ exponential in the mixing time $\bar{b}$, which is large exactly when some states are hard to reach. *(ii) Bias at skipped states.* A rollout that stops once the active set is covered leaves the estimates at the other states uncontrolled, and the convergence proof must absorb this bias. Last-iterate analyses that compare the current iterate with earlier ones need bias control along those comparisons, which these estimates do not provide. *(iii) Resolution without observed gaps.* The active set is read off the action probabilities, not the unknown gaps. Through the value error, one can certify that a state with $\Delta(s) > 0$ has left it only once the value error is below $\underline{\pi}\,\Delta(s)$, which adds a factor $\underline{\pi}^{-2}$ to the exploration bound. *(iv) A lower bound in the same model.* A matching lower bound must hold on action-dependent dynamics and within the model of the upper bound. Its strongest form uses deterministic costs, which reveal any gap in immediate costs after one observation, so the small gaps must come from transitions.

Contributions. *Conceptually*, we identify the object that controls the exploration cost: the *certified hitting cost* $H_\theta(s)$, equivalent up to constants to the worst-case expected time to reach s under the *certified* policies, those that keep mass at least $\theta := \underline{\pi}$ on optimal actions (Definition 3.1, Proposition 3.2). With high probability every SPMD iterate is certified, and H_θ involves no stationary distribution and no mixing time. Exploration is charged as reachability times resolution scale, and only while a state is unresolved:

$$\mathcal{I}_\infty(H_\theta;\varepsilon) = \max_{s\in\mathcal{S}} \frac{H_\theta(s)}{\max\{\varepsilon,\Delta(s)\}^2}.$$

Technically, we close the four obstacles in turn. (i) Because every iterate is certified, a pair-hitting bound uniform over the certified class controls every rollout through H_θ, with no mixing parameter of intermediate policies (Lemma C.1). H_θ itself is bounded through first-passage comparison with the optimal chain (Lemma 3.9) or by hitting estimates uniform over policies (Lemma 4.2). (ii) The truncated estimates never favor a suboptimal action over the optimal one by more than the bias tolerance, and a last-iterate proof in cumulative-gradient form uses only comparisons with the deterministic optimal policy (Lemma 5.2, Section 5.1). (iii) The same structure bounds the advantage of every non-rare action through its dual coordinate, so the exploration bound avoids the factor $\underline{\pi}^{-2}$ of obstacle (iii) (Lemma 5.3). (iv) A ring whose small gaps come from transitions supports a lower bound with deterministic costs (Section 4).

Formally:

- Theorem 3.7: if some deterministic optimal policy induces an irreducible chain, SPMD with an active-set-aware stopping rule uses $K(1+L_\gamma) + \widetilde{O}(|\mathcal{A}|^{1-p}\theta^{-1}(1-\gamma)^{-4}\,\mathcal{I}_\infty(H_\theta;(1-\gamma)\varepsilon))$ samples, where the baseline $K(1+L_\gamma) = \widetilde{O}(|\mathcal{A}|^{1-p}(1-\gamma)^{-7}\varepsilon^{-2})$ is instance-independent.
- Corollaries 3.11–3.12: under geometric ergodicity, $H_\theta(s) \le 2\bar{b}/(\theta^{\bar{b}}\nu^*(s))$, which, unlike the analysis of Ju & Lan (2025), does not multiply separate bounds on the mixing and the stationary mass of intermediate policies; under action-independent transitions, H_θ is the reachability cost of the common chain.
- Section 4: on an action-dependent ring with $3n-2$ states, the exploration term is $\widetilde{O}(n^2 + n\varepsilon^{-2})$ and every algorithm needs $\Omega(\max\{n^2, n\varepsilon^{-2}\})$ samples, both with deterministic costs; charging the hardest state at every iteration, at the same hitting cost, gives $\Theta(n^2\varepsilon^{-2})$.

Scope. The results are instance-dependent, not instance-optimal: the lower bound matches on one family, in (n,ε) at fixed γ and $|\mathcal{A}|$, and the dependence on γ and $|\mathcal{A}|$ is not claimed optimal (Remark 3.8).

2 Related work and preliminaries

Auto-exploration and online discounted RL. SPMD (Lan, 2023) is a policy gradient method based on mirror descent on the action simplex. Li & Lan (2025) discovered that SPMD with Tsallis entropy exhibits *implicit exploration*: the policy automatically maintains sufficient support over actions without explicit ε-greedy perturbation, yielding an $O(\varepsilon^{-2})$ sample complexity whose constants depend on the iterates. Ju & Lan (2025) then strengthened this to an *algorithm-independent* result by introducing a dynamic mixing time that adapts to the chain's behavior without prior knowledge of

mixing parameters. Nanda & Chen (2025) established similar guarantees for on-policy Q-learning under minimal assumptions, using a Poisson-equation decomposition to handle time-varying Markov noise. Table 1 summarizes how our work differs from Ju & Lan (2025).

Table 1: Comparison with the original auto-exploration paper. H_θ is the certified hitting cost (Definition 3.1) and $\mathcal{I}_\infty = \mathcal{I}_\infty(H_\theta; (1-\gamma)\varepsilon)$ the local exploration complexity (Definition 3.4).

	Ju & Lan (2025)	This paper
Algorithm	SPMD + dynamic mixing	Same + active-set-aware mixing
Assumption on π^*	Irreducible, geometrically ergodic	Irreducible
Exploration complexity	Global D_{expl}	Local $\mathcal{I}_\infty(H_\theta; (1-\gamma)\varepsilon)$
Theorem shape	$\widetilde{O}(D_{\text{expl}} \cdot \varepsilon^{-2})$	$K(1+L_\gamma) + \widetilde{O}\big(\frac{\lvert\mathcal{A}\rvert^{1-p}}{\underline{\pi}(1-\gamma)^4}\mathcal{I}_\infty\big)$
Lower bound	None	Matching on an action-dependent family

Gap dependence, instance dependence and lower bounds. Policy gradient and policy mirror descent converge linearly when all advantage gaps are bounded away from zero (Xiao, 2022; Li et al., 2024); these rates describe optimization speed, not the sample cost of trajectory-based exploration. Our locking argument (Lemma 5.3) instead works through dual coordinates to shrink the active set. Instance-dependent regret and PAC guarantees, including the interplay of reachability and gaps, are known for episodic MDPs (Simchowitz & Jamieson, 2019; Wagenmaker et al., 2022; Tirinzoni et al., 2022), and matching lower bounds are classical with a generative model (Gheshlaghi Azar et al., 2013; Agarwal et al., 2020). The online single-trajectory discounted setting has no resets and no simulator. To our knowledge, prior instance-dependent exploration bounds for it hold only in the asymptotic regime $\delta \to 0$ (Al Marjani et al., 2021); Appendix M discusses these connections.

MDP and SPMD. We consider an infinite-horizon discounted MDP $(\mathcal{S}, \mathcal{A}, P, c, \gamma)$ with finite state and action spaces, transition kernel $P(s'|s,a)$, costs $c : \mathcal{S}\times\mathcal{A} \to [0,1]$, and discount factor $\gamma \in [0,1)$. The agent observes costs $C_t \in [0,1]$ with $\mathbb{E}[C_t \mid \mathcal{H}_t] = c(S_t, A_t)$, where $\mathcal{H}_t$ is the history up to (S_t, A_t), and $S_{t+1} \sim P(\cdot|S_t, A_t)$ given $(\mathcal{H}_t, C_t)$. The upper bounds hold in this model; the lower bounds (Section 4.2) use deterministic costs $C_t = c(S_t, A_t)$. Reward maximization reduces to this model through $C_t = 1 - R_t$ (Remark E.1). For any stationary policy π, write $P_\pi^{(t)}(s_0, \cdot)$ for the distribution of the state after t steps starting from s_0. The value function and Q-function under policy π are $V^\pi(s) := \mathbb{E}_\pi[\sum_{t=0}^\infty \gamma^t c(s_t,a_t) \mid s_0 = s]$ and $Q^\pi(s,a) := c(s,a) + \gamma \sum_{s'} P(s'|s,a)V^\pi(s')$. The optimal policy π^* minimizes $V^\pi(s)$. As in Ju & Lan (2025), SPMD uses the Tsallis-p Bregman divergence $D_x^y := \frac{1}{p(1-p)}\sum_a[(1-p)x_a^p - y_a^p + p\, y_a x_a^{p-1}] \geq 0$, the Bregman divergence of y from the base point x generated by the convex function $-\sum_a v_a^p/(p(1-p))$. We focus on the unregularized case, with no convex regularizer h added to the cost ($h = 0$).

Each iteration t runs the chain for a *rollout* to collect an estimate $\tilde{Q}^{\pi_t}$ and then takes a mirror-descent step; the rollout length, the *dynamic mixing time*, is the object we modify.

Assumption 2.1 (Irreducible optimal policy)**.** There exists a deterministic optimal policy π^* whose induced chain P_{π^*} is irreducible.

We fix one such π^* and write $a^*(s)$ for the action it selects at s. When several actions are optimal at s, this choice is the tie-breaking rule, and all instance-dependent quantities below are defined relative to it. Assumption 2.1 is the only structural assumption of the main theorem. Geometric ergodicity, which Ju & Lan (2025) assume throughout, enters only when we compare with their mixing-based analysis (Assumption 3.10 and Corollary 3.11).

The TOMC estimator. To estimate Q-values from a single trajectory, Ju & Lan (2025) use a truncated on-policy Monte Carlo (TOMC) estimator: run the chain for m steps and estimate each visited pair by its discounted return from the first visit. Pairs of probability below the threshold $\underline{\pi}$ (*rare* pairs) may never be visited, and their estimate is set pessimistically to $(1-\gamma)^{-1}$; $\underline{\pi}$ is fixed in Section 3. Formally, denote $\mathcal{Z} := \mathcal{S}\times\mathcal{A}$. The set of *non-rare* pairs under policy π is $\mathcal{Z}_{\text{nr}}^\pi(\underline{\pi}) := \{(s,a) \in \mathcal{Z} : \pi(a|s) \geq \underline{\pi}\}$. Given a rollout of length m, the TOMC estimator is

$$\tilde{Q}_{m,\underline{\pi}}(z) := \begin{cases} \displaystyle\sum_{t=\tau_m^\pi(z)}^{m-1} \gamma^{t-\tau_m^\pi(z)}\, C_t & z \in \mathcal{Z}_{\text{nr}}^\pi(\underline{\pi}), \\ (1-\gamma)^{-1} & z \notin \mathcal{Z}_{\text{nr}}^\pi(\underline{\pi}), \end{cases} \tag{1}$$

where $Z_t = (s_t, a_t)$ is the state-action pair at step t of the rollout, C_t the observed cost, and $\tau_m^\pi(z) := \min\{t \in [0, m) : Z_t = z\}$ is the first hitting time of pair z (set to m if z is not visited). We write $\mathcal{F}_t$ for the σ-algebra generated by all randomness before the rollout of iteration t.

Dynamic mixing time. The rollout length m is not fixed in advance but chosen adaptively. Define the *discounted hitting residual* $h_m^\pi(z) := \gamma^{m-\tau_m^\pi(z)}$, which measures how much discounted weight remains after hitting pair z: when $h_m^\pi(z)$ is small, the truncation bias at z is small. For a bias tolerance ς, the *dynamic mixing time* $\tilde{m}^\pi(\underline{\pi}, \varsigma)$ is the smallest m such that $\max_{z \in \mathcal{Z}_{\mathrm{nr}}^\pi(\underline{\pi})} h_m^\pi(z) \le \varsigma(1-\gamma)$, ensuring bias $\le \varsigma$ at all non-rare pairs (see Ju & Lan, 2025, Lemma 3.1).

3 Main results

3.1 Active-set-aware mixing

We introduce a single modification to the dynamic mixing time of Ju & Lan (2025). Define the *multi-action states*, which form the *active set* of iteration t:

$$\mathcal{S}_t^{\mathrm{multi}} := \{s \in \mathcal{S} : |\{a : \pi_t(a|s) \ge \underline{\pi}\}| \ge 2\}. \tag{2}$$

These are states where the current policy π_t has not yet concentrated on a single action. The *localized dynamic mixing time* restricts bias control to these states: informally, the rollout runs until every non-rare pair at a multi-action state has been visited with sufficient discounted weight. Formally:

$$\tilde{m}_t^{\mathrm{loc}}(\underline{\pi}, \varsigma) := \min_{m \ge 1} \Big\{ m : \max_{\substack{z=(s,a) \in \mathcal{Z}_{\mathrm{nr}}^{\pi_t}(\underline{\pi}) \\ s \in \mathcal{S}_t^{\mathrm{multi}}}} h_m^{\pi_t}(z) \le \varsigma(1-\gamma) \Big\}. \tag{3}$$

When $\mathcal{S}_t^{\mathrm{multi}} = \varnothing$, $\tilde{m}_t^{\mathrm{loc}} = 1$.

Parameter setting. We fix the Tsallis entropic index $p := 1/\log_2((1-\gamma)^{-1})$ and write $\bar{Q} := (1-\gamma)^{-1}$ and $D_0 := (|\mathcal{A}|^{1-p} - 1)/(p(1-p))$, the divergence of every deterministic policy row from the uniform one. For target accuracy ε and failure probability δ, we run $K = \widetilde{\Theta}(|\mathcal{A}|^{1-p}(1-\gamma)^{-6}\varepsilon^{-2})$ iterations (explicit choice in Appendix G) with stepsizes $\eta_t = \alpha/\sqrt{t+1}$, $\alpha := \sqrt{D_0 + 1}/(\bar{Q}\sqrt{\ln(eK)})$. We set the rare-action threshold $\underline{\pi}$ as in Lemma G.8 and the bias tolerance $\varsigma := (1-\gamma)^2\varepsilon/2$ in (3). Logarithmic factors enter only through K and the confidence budgets; no parameter depends on reachability or mixing quantities.

Algorithm 1 SPMD with active-set-aware mixing

1: Initialize $\pi_0(\cdot|s) \leftarrow \mathrm{Uniform}(\mathcal{A})$ for all s
2: **Set stepsizes** $\eta_t \leftarrow \alpha/\sqrt{t+1}$ for $t = 0, \dots, K-1$
3: **for** $t = 0, \dots, K-1$ **do**
4: **Compute** $\tilde{m}_t^{\mathrm{loc}}$ via (3) ▷ See below for two implementations
5: Run TOMC estimator with mixing time $\tilde{m}_t^{\mathrm{loc}}$ to get $\tilde{Q}^{\pi_t}$
6: Update $\pi_{t+1}(\cdot|s) \leftarrow \mathrm{argmin}_{\upsilon \in \Delta_{|\mathcal{A}|}} \{\langle \tilde{Q}^{\pi_t}(s, \cdot), \upsilon \rangle + \eta_t^{-1} D_{\pi_t}^{\upsilon}(s)\}$
7: **end for**
8: **return** π_K

Implementation. Compute $\mathcal{S}_t^{\mathrm{multi}}$ by scanning policy rows ($O(|\mathcal{S}||\mathcal{A}|)$), then run the rollout until $\max_{(s,a) \in \mathcal{Z}_{\mathrm{nr}}^{\pi_t}(\underline{\pi}),\, s \in \mathcal{S}_t^{\mathrm{multi}}} h_m^{\pi_t}(s,a) \le \varsigma(1-\gamma)$. An equivalent on-the-fly variant (Proposition B.1) classifies each state lazily upon its first visit in the rollout and maintains a running set of pairs not yet covered, avoiding the upfront $O(|\mathcal{S}||\mathcal{A}|)$ scan.

Why skipping single-action states is safe. The algorithm does not observe $\Delta(s)$. Once the locking scale of Lemma 5.3 drops below $\Delta(s)$, every suboptimal action at s is rare, so s leaves $\mathcal{S}_t^{\mathrm{multi}}$; it never returns, because rare actions stay rare under the mirror step (Lemma B.2). Thus $\mathcal{S}_t^{\mathrm{multi}}$ is a computable, shrinking proxy for the set of unresolved states.

3.2 Certified hitting cost

The exploration cost of an iteration is the time the rollout needs to visit every non-rare pair at a multi-action state, and it depends on the current policy π_t, which is random. An algorithm-independent

bound therefore needs a class of policies that provably contains every iterate and over which hitting times are controlled uniformly. For a support level $\theta \in (0, 1]$, the class of *certified* policies is

$$\Pi_\theta := \big\{\pi \text{ stationary} : \ \pi(a^*(s)|s) \geq \theta \ \ \forall s \in \mathcal{S}\big\}.$$

Algorithm 1 takes $\theta = \underline{\pi}$; with high probability every iterate then lies in Π_θ (Section 5.1). For $s \in \mathcal{S}$ let $\tau_s := \min\{t \geq 0 : s_t = s\}$.

Definition 3.1 (Certified hitting cost)**.** For $s \in \mathcal{S}$ and $\ell \geq 1$, let $q_s^\theta(\ell) := \inf_{\pi \in \Pi_\theta} \inf_{x \in \mathcal{S}} \Pr_x^\pi(\tau_s \leq \ell)$. The *certified hitting cost* of s is $H_\theta(s) := \inf_{\ell \geq 1} \ell / q_s^\theta(\ell)$.

$H_\theta(s)$ is a worst case over starting states and certified policies and involves nothing else: no stationary distribution, no mixing time, and no policy outside Π_θ. It is equivalent up to constants to a standard hitting time, and it is finite exactly under Assumption 2.1.

Proposition 3.2 (H_θ as a hitting time)**.** *Let $D_\theta(s) := \sup_{\pi \in \Pi_\theta} \max_{x \in \mathcal{S}} \mathbb{E}_x^\pi[\tau_s]$. Then $D_\theta(s) \leq H_\theta(s) \leq 4D_\theta(s) + 2$ for every s. Moreover, $H_\theta(s) < \infty$ for all s if and only if P_{π^*} is irreducible.*

3.3 Local exploration complexity

Definition 3.3 (Statewise advantage gap)**.** $\Delta(s) := \min_{a \neq a^*(s)}[Q^{\pi^*}(s, a) - V^{\pi^*}(s)] \geq 0$. When $\Delta(s) = 0$ (ties), the analysis treats s as unresolved at every accuracy level; the algorithm may or may not keep such a state in $\mathcal{S}_t^{\text{multi}}$.

Definition 3.4 (Local exploration complexity)**.** For a state weight $w : \mathcal{S} \to [0, \infty)$ and accuracy $\varepsilon \in (0, 1)$, let

$$\mathcal{I}_\infty(w; \varepsilon) := \max_{s \in \mathcal{S}} \frac{w(s)}{\max\{\varepsilon, \Delta(s)\}^2}.$$

The complexity governing Theorem 3.7 is $\mathcal{I}_\infty(H_\theta; (1 - \gamma)\varepsilon)$.

Each state is charged its reachability cost at its own resolution scale: a state with $\Delta(s) > \varepsilon$ is charged for the $\Delta(s)^{-2}$ rounds needed to resolve it, and only states with $\Delta(s) \leq \varepsilon$ are charged for the full ε^{-2}.

The proof tracks a finer, phase-by-phase version. With $J := \lceil \log_2(1/\varepsilon) \rceil$, $j_{\min} := -\lceil \log_2((1 - \gamma)^{-1}) \rceil$ and $\max \varnothing := 0$, let

$$\mathcal{I}_{\text{ph}}(w; \varepsilon) := \sum_{j=j_{\min}}^{J} 4^j \max_{s:\, \Delta(s) \leq 2^{-j}} w(s). \tag{4}$$

In phase j the accuracy is $\asymp 2^{-j}$, the phase lasts $\asymp 4^j$ units, and only states with $\Delta(s) \leq 2^{-j}$ are unresolved. The sum starts at $j_{\min}$ so that the first phase contains every state, since $\Delta(s) \leq (1 - \gamma)^{-1} \leq 2^{-j_{\min}}$.

Proposition 3.5 (Single-state form)**.** *For every weight w, $\frac{1}{4}\mathcal{I}_\infty(w; \varepsilon) \leq \mathcal{I}_{\text{ph}}(w; \varepsilon) \leq 4\,(J - j_{\min} + 1)\,\mathcal{I}_\infty(w; \varepsilon)$. The factor $J - j_{\min} + 1$ cannot be replaced by a constant.*

Up to one logarithmic factor, the exploration term is therefore controlled by the largest product of a state's reachability cost and its resolution scale. The quantity $\mathcal{I}_{\text{ph}}$ keeps the phase-by-phase accounting that the proof uses, in which different phases can be dominated by different states. Both are dominated by the charge that bills the hardest state at every phase.

Proposition 3.6 (Local versus global charge)**.** *For every weight w, $\mathcal{I}_\infty(w; \varepsilon) \leq \varepsilon^{-2} \max_s w(s)$ and $\mathcal{I}_{\text{ph}}(w; \varepsilon) \leq \sum_{j=j_{\min}}^{J} 4^j \max_s w(s) \leq \frac{16}{3}\varepsilon^{-2} \max_s w(s)$.*

Both inequalities follow from $\max_{\Delta(s) \leq 2^{-j}} w(s) \leq \max_s w(s)$ and $4^J < 4\varepsilon^{-2}$. They are tight up to constants when a maximizer of w has $\Delta(s) \leq \varepsilon$, and loose by a factor $\Theta(\min\{n, \varepsilon^{-2}\})$ on the family of Section 4.

3.4 Sample complexity bound

The theorem splits the sample count into an instance-independent baseline (optimization plus residual decay) and an exploration term governed by $\mathcal{I}_{\text{ph}}(H_\theta; (1 - \gamma)\varepsilon)$.

Theorem 3.7 (Instance-dependent auto-exploration)**.** *Fix $\varepsilon, \delta \in (0,1)$ and $\gamma \in [3/4, 1)$. Under Assumption 2.1, run Algorithm 1 with $h = 0$ and the parameters of Section 3, and let*

$$L_\gamma(\varsigma) := \left\lceil \frac{\log(1/(\varsigma(1-\gamma)))}{|\log\gamma|} \right\rceil.$$

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

(Convergence.) *$V^{\pi_K}(s) - V^{\pi^*}(s) \le \varepsilon$ for all $s \in \mathcal{S}$.*

(Sample complexity.)

$$\begin{aligned} \textit{total samples} \;&\le\; \underbrace{K\big(1 + L_\gamma(\varsigma)\big)}_{\textit{baseline}} \;+\; \underbrace{\widetilde{O}\bigg(\frac{|\mathcal{A}|^{1-p}}{\theta\,(1-\gamma)^4}\,\mathcal{I}_{\mathrm{ph}}\big(H_\theta; (1-\gamma)\varepsilon\big)\bigg)}_{\textit{exploration}} \\ &\le\; K\big(1 + L_\gamma(\varsigma)\big) \;+\; \widetilde{O}\bigg(\frac{|\mathcal{A}|^{1-p}}{\theta\,(1-\gamma)^4}\,\mathcal{I}_{\infty}\big(H_\theta; (1-\gamma)\varepsilon\big)\bigg). \end{aligned} \tag{5}$$

Reading the bound. The baseline is $\widetilde{O}(|\mathcal{A}|^{1-p}(1-\gamma)^{-7}\varepsilon^{-2})$ and does not depend on the instance. The exploration term charges each state its certified hitting cost at resolution $\max\{(1-\gamma)\varepsilon, \Delta(s)\}^{-2}$. The resolution is $(1-\gamma)\varepsilon$ because states lock through their one-step advantages, which must be accurate to about $(1-\gamma)\varepsilon$ for an ε-accurate value. The single factor θ^{-1} comes from selecting a non-rare action once its state is reached (Lemma C.1), and locking costs no power of θ (Lemma 5.3). All instance dependence sits in H_θ and the gaps.

Remark 3.8 (Dependence on γ and $|\mathcal{A}|$)**.** The iteration count is $K = \widetilde{\Theta}(|\mathcal{A}|^{1-p}(1-\gamma)^{-6}\varepsilon^{-2})$. The last-iterate analysis of Ju & Lan (2025) (Lemmas A.1–A.2 of its arXiv version 4) would give $(1-\gamma)^{-4}$, but its proof needs bias control along the direction of earlier iterates, which the localized estimator does not provide (Section 5.1). The direct last-iterate bound costs $(1-\gamma)^{-2}$ in the baseline. In the exploration term, locking uses the one-step advantage, a factor $1-\gamma$ smaller than the value error (Lemma 5.3). This gives the prefactor $(1-\gamma)^{-4}$, and the remaining loss appears only through the resolution $(1-\gamma)\varepsilon$, i.e., for states with $\Delta(s) < \varepsilon$. Since $\theta = \widetilde{\Theta}((1-\gamma)/|\mathcal{A}|)$ (Lemma G.8), the explicit prefactor outside the complexity is $\widetilde{O}(|\mathcal{A}|^{2-p}(1-\gamma)^{-5})$; further dependence on γ may enter through H_θ, the gaps and the resolution. We make no optimality claim in γ or $|\mathcal{A}|$.

3.5 Evaluating the certified hitting cost

One way to evaluate H_θ is to compare a certified policy with π^* along first-passage paths. Let $\mathcal{S}_{\mathrm{dep}}$ be the set of states whose transition kernel depends on the action, and $N_\tau := |\{k < \tau : s_k \in \mathcal{S}_{\mathrm{dep}}\}|$.

Lemma 3.9 (Path comparison)**.** *For every $\pi \in \Pi_\theta$, $x, s \in \mathcal{S}$ and $\ell \ge 1$,*

$$\Pr_x^{\pi}(\tau_s \le \ell) \;\ge\; \mathbb{E}_x^{\pi^*}\big[\theta^{N_{\tau_s}} \mathbf{1}\{\tau_s \le \ell\}\big] \;\ge\; \theta^\ell \Pr_x^{\pi^*}(\tau_s \le \ell).$$

A step from a state whose kernel does not depend on the action is free; only steps through $\mathcal{S}_{\mathrm{dep}}$ pay θ. The fixed-time comparison $P_\pi^{(\ell)}(x,s) \ge \theta^\ell P_{\pi^*}^{(\ell)}(x,s)$ of Ju & Lan (2025) is the special case of the event $\{s_\ell = s\}$ with every step charged. The comparison gives the mixing-based estimate below and the example that closes this section; the ring of Section 4 instead admits hitting bounds that hold uniformly over all policies (Lemma 4.2). The mixing-based estimate uses the assumption of Ju & Lan (2025).

Assumption 3.10 (Geometric ergodicity)**.** The chain P_{π^*} has stationary distribution ν^* with $\nu^*(s) \ge \underline{\nu}^* > 0$, and $\|P_{\pi^*}^{(t)}(s_0, \cdot) - \nu^*\|_{\mathrm{tv}} \le C^*(\rho^*)^t$ for all s_0 and t, with $C^* \le 2$ and $\rho^* \in [1/2, 1)$.

Corollary 3.11 (Geometrically ergodic π^*)**.** *Under Assumption 3.10, let $\bar{b} := \lceil \log(4/\underline{\nu}^*)/\log(1/\rho^*) \rceil$. Then $H_\theta(s) \le 2\bar{b}/(\theta^{\bar{b}}\nu^*(s))$ for every s, and the exploration term of Theorem 3.7 is at most $\widetilde{O}\big(|\mathcal{A}|^{1-p}\,\bar{b}\,\theta^{-(\bar{b}+1)}(1-\gamma)^{-4}\,\mathcal{I}_\infty(1/\nu^*; (1-\gamma)\varepsilon)\big)$.*

This is the regime of Ju & Lan (2025). Their conversion multiplies separate lower bounds on the mixing rate and on the stationary mass of every intermediate policy (Ju & Lan, 2025, Eq. (26));

hitting the target directly avoids this product, and Remark C.2 compares the explicit factors. A single factor $\theta^{-\bar{b}}$ remains, and it is exponential in the mixing time: $\theta^{-\bar{b}} \le \theta^{-1}(4/\underline{\nu}^*)^{\log(1/\theta)/\log(1/\rho^*)}$. On slowly mixing instances this regime is uninformative, which is why Section 4 evaluates H_θ directly. At the other extreme, the dynamics do not depend on the action.

Corollary 3.12 (Policy-invariant dynamics)**.** *If $P(\cdot|s,a)$ does not depend on a, then for every $\theta \in (0,1]$, $H_\theta(s) = H_{\mathrm{reach}}(s) := \inf_{\ell \ge 1} \ell / \inf_x \Pr_x(\tau_s \le \ell)$, the reachability cost of the common chain.*

Between these extremes, only visits to $\mathcal{S}_{\mathrm{dep}}$ are charged. On the action-dependent star of Appendix J, where $n-1$ states surround a hub h and only the bottleneck is action-dependent, Lemma 3.9 with $\ell = 2$ gives $H_\theta(s) \le 2/(\theta P(h,s))$: a fixed power of θ, where Corollary 3.11 would pay $\theta^{-\Theta(\log n)}$.

4 Sharpness on an action-dependent family

The family below separates reachability from resolution: one state is hard to reach but easy to resolve, while the others are easy to reach but hard to resolve. Two further properties matter for the lower bound. The transitions depend on the action at the branch states, which stay unresolved longest, and the small gaps come from transitions, so both bounds hold with deterministic costs.

4.1 A transition-gap ring

Example 4.1 (Transition-gap ring)**.** Fix $n \ge 4$, $r := n^{-2}$, $m := n-1$ and $\eta \in (0, 1/4]$. The $N = 3n-2$ states are a bottleneck b and, for $i \in [m]$, a branch state g_i with successors x_i, y_i. Costs are deterministic and depend only on the state: $c(b) = c(x_i) = 1$ and $c(g_i) = c(y_i) = 0$. Actions are $\{L, R\}$. From every state other than b, the chain jumps to b with probability r under either action. Otherwise, from g_i it moves to y_i with probability $\frac{1}{2} + \zeta_a$ and to x_i with probability $\frac{1}{2} - \zeta_a$, where $\zeta_L := \eta$ and $\zeta_R := -\eta$; from x_i and y_i it moves to g_{i+1} under either action ($g_{m+1} := g_1$). From b, action L moves to g_1, and action R stays at b or moves to g_1 with probability $\frac{1}{2}$ each.

Lemma 4.2 (Parameters of the ring)**.** *Let $\gamma \in [3/4, 1)$ and $a_r := 1 - r$.*

(i) The policy that plays L everywhere is optimal, with $\Delta(g_i) = 2\gamma a_r \eta$, $\frac{45}{256} \le \frac{\gamma a_r}{4} \le \Delta(b) \le \frac{\gamma}{2}$, and $\Delta(x_i) = \Delta(y_i) = 0$ (the two actions coincide at x_i, y_i).
(ii) Assumptions 2.1 and 3.10 hold, with $\nu^*(b) = r/(1+r)$, $\nu^*(s) = \Theta(1/n)$ for $s \ne b$, $C^* = 2$ and $\rho^* = \sqrt{1 - r a_r} = 1 - \Theta(n^{-2})$; hence $\underline{\nu}^* = \Theta(n^{-2})$ and $\bar{b} = \Theta(n^2 \log n)$.
(iii) For every $\theta \in (0,1]$: $H_\theta(b) = n^2$, and $n/2 \le H_\theta(s) \le 64n/3$ for every $s \ne b$.

The bottleneck is hard to reach but has a constant gap; every other state is reached within $O(n)$ steps, and the branch states have gap $\Theta(\eta)$. The gaps at g_i come from transitions: the two actions share the cost and the jump probability and differ only in how the remaining mass is split, while $V^\pi(x_i) - V^\pi(y_i) = 1$ for every policy π. Resolving g_i requires learning a transition probability, not comparing immediate costs. The chain mixes only through resets at b, so $\bar{b} = \Theta(n^2 \log n)$ and the mixing regime (Corollary 3.11) pays $\theta^{-\Theta(n^2 \log n)}$ here; the bound below comes from H_θ directly.

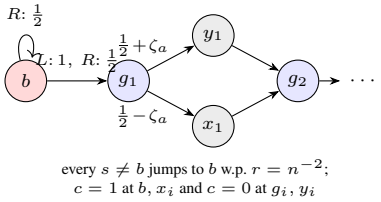


Figure 1: The transition-gap ring. The two actions at g_i have the same cost and the same jump probability; they differ only in the split between x_i and y_i, whose values differ by exactly 1.

Proposition 4.3 (Upper bound on the ring)**.** *On Example 4.1 with $\gamma \in [3/4, 1)$ fixed and any $\eta \in (0, 1/4]$, the exploration term of Theorem 3.7 is $\widetilde{O}(n^2 + n\varepsilon^{-2})$ for every $\varepsilon \in (0,1)$.*

Separation. At the same hitting cost, the charge that bills the hardest state at every phase is $\varepsilon^{-2} \max_s H_\theta(s) = \Theta(n^2 \varepsilon^{-2})$ (Proposition 3.6), against $\Theta(n^2 + n\varepsilon^{-2})$ here: the ratio is $\Theta(\min\{n, \varepsilon^{-2}\})$. This is the gain from charging only unresolved states. It is separate from the gain of hitting targets directly rather than through mixing, which on this family is the difference between $\theta^{-\Theta(n^2 \log n)}$ and a fixed power of θ. Appendix A runs Algorithm 1 on the ring: once b is resolved, local rollouts are about 16 times shorter, and the end-to-end saving grows with the target

accuracy, to about 3 times at error 10^{-3}. Remark K.1 in Appendix K explains the role of the branch layer and what the family does not show.

4.2 A MATCHING LOWER BOUND WITH DETERMINISTIC COSTS

The lower bound uses a uniform PAC criterion. An algorithm interacts with an instance from a finite family for T transitions, starting at g_2, and must output a stationary policy $\hat{\pi}$ with $V^{\hat{\pi}}(s) - V^{\pi^*}(s) \le \varepsilon$ for all s, with probability at least $2/3$, on every instance of the family. Uniformity is necessary: an algorithm tuned to a single known instance can output its optimal policy without sampling. Each family below consists of Example 4.1, denoted $\mathcal{M}_0$, and one relabeled copy $\mathcal{M}_1$; costs are deterministic in both.

Proposition 4.4 (Bottleneck lower bound). *Let $\mathcal{M}_1$ be $\mathcal{M}_0$ with the kernels of L and R at b exchanged. If $\gamma \in [3/4, 1)$ and $\varepsilon < 1/12$, any algorithm meeting the criterion on $\{\mathcal{M}_0, \mathcal{M}_1\}$ uses $T \ge n^2/3$ transitions.*

Proposition 4.5 (Branch-state lower bound). *Choose η so that $\Delta(g_1) = 3\varepsilon$, and let $\mathcal{M}_1$ be $\mathcal{M}_0$ with ζ_L and ζ_R exchanged at g_1 only. If $\gamma \in [3/4, 1)$ and $\varepsilon < 1/12$, then $\eta \le \frac{32}{15}\varepsilon < \frac{1}{4}$, and any algorithm meeting the criterion on $\{\mathcal{M}_0, \mathcal{M}_1\}$ uses $T \ge n/(90\eta^2) = \Omega(n\varepsilon^{-2})$ transitions.*

The bottleneck bound needs no divergence: the two instances coincide until b is first visited, which by time T happens with probability at most Tr. The branch-state bound combines the occupancy bound $\mathbb{E}[N_{g_1}(T)] \le 5T/(4n)$, valid for every adaptive algorithm, with the divergence decomposition for instances that differ in their transitions (Appendix I).

Corollary 4.6 (Matching bounds). *Fix $\gamma \in [3/4, 1)$, $n \ge 4$ and $\varepsilon \in (0, 1/12)$, let $\eta := 3\varepsilon/(2\gamma(1-r))$, and let $\mathcal{F}$ consist of Example 4.1 and its two relabelings from Propositions 4.4 and 4.5. Every algorithm meeting the uniform criterion on $\mathcal{F}$ requires $\Omega(\max\{n^2, n\varepsilon^{-2}\})$ transitions, and Algorithm 1 uses $\widetilde{O}(n^2 + n\varepsilon^{-2})$ transitions on every instance of $\mathcal{F}$ (the baseline, $\widetilde{O}(\varepsilon^{-2})$ for fixed γ, is absorbed). The two agree up to logarithmic factors, with deterministic costs on both sides.*

5 PROOF SKETCH

Complete proofs are in the appendix. Section 5.1 proves convergence despite the bias at skipped states, obstacle (ii), and certifies every iterate; this is Part A. Section 5.2 shows that a state leaves the active set once it is resolved, obstacle (iii), and bounds the rollouts through certified pair hitting, obstacle (i); this is Part B.

5.1 RELATIVE BIAS AND DIRECT LAST-ITERATE CONVERGENCE

Let $b_t(s,a) := \mathbb{E}[\tilde{Q}^{\pi_t}(s,a) \mid \mathcal{F}_t] - Q^{\pi_t}(s,a)$ denote the conditional bias. The localized rollout controls b_t only at non-rare pairs of multi-action states; elsewhere the truncated estimate can miss most of the return. What survives everywhere is a sign structure relative to the optimal action.

Lemma 5.1 (One-sided TOMC bias). *For every non-rare pair $z \in \mathcal{Z}^{\pi_t}_{\mathrm{nr}}(\underline{\pi})$, $\mathbb{E}[\tilde{Q}^{\pi_t}(z) \mid \mathcal{F}_t] \le Q^{\pi_t}(z)$.*

Lemma 5.2 (Relative bias). *Fix t and condition on $\mathcal{F}_t$. Suppose no rare action is the action selected by π^*, i.e., $\pi_t(a|s) < \underline{\pi} \Rightarrow \pi^*(a|s) = 0$. Then $b_t(s,a) - b_t(s, a^*(s)) \ge -\varsigma$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. Consequently $\langle b_t(s,\cdot), v - \pi^*(\cdot|s)\rangle \ge -\varsigma$ for every $v \in \Delta_{|\mathcal{A}|}$; in particular the weighted bias $\langle -b_t(s,\cdot), \pi_t(\cdot|s) - \pi^*(\cdot|s)\rangle$ is at most ς at multi-action states and at most 0 at single-action states.*

Three signs combine: non-rare estimates are truncations of a non-negative return ($b_t \le 0$, Lemma 5.1); non-rare pairs at multi-action states are estimated to within ς by the stopping rule ($b_t \ge -\varsigma$); rare actions receive the pessimistic default ($b_t \ge 0$).

The quantifier over v is what the convergence proof uses. Last-iterate analyses that compare the current iterate with earlier ones (Jain et al., 2021; Ju & Lan, 2025) need the bias along $\pi_t - \pi_{t_0}$, which Lemma 5.2 does not control: a suboptimal action that has become rare carries the pessimistic default and may have more mass under π_{t_0}. Part A therefore compares only with π^*.

For $h = 0$ the Tsallis iterates stay in the interior of the simplex, and summing the optimality conditions of the mirror steps gives

$$\pi_t(\cdot|s) = \arg\min_{v \in \Delta_{|\mathcal{A}|}} \Big\{ \Big\langle \sum_{u<t} \eta_u \tilde{Q}^{\pi_u}(s,\cdot),\, v \Big\rangle + D^v_{\pi_0}(s) \Big\}.$$

Testing against $v = \pi^*(\cdot|s)$ bounds $\langle \bar{Q}_t(s,\cdot), \pi_t(\cdot|s) - \pi^*(\cdot|s)\rangle$, where $\bar{Q}_t$ is the stepsize-weighted average of Q^{π_u} over $u < t$. The bias enters only as $\sum_{u<t} \eta_u \langle b_u(s,\cdot), \pi_t(\cdot|s) - \pi^*(\cdot|s)\rangle$, which Lemma 5.2 bounds pathwise, and the noise is controlled coordinatewise. Since $\bar{Q}_t = c + \gamma P \bar{V}_t$ with $\bar{V}_t$ the average of V^{π_u}, this gives $\|V^{\pi_t} - V^{\pi^*}\|_\infty \le \varepsilon_t = \widetilde{O}(|\mathcal{A}|^{(1-p)/2}(1-\gamma)^{-3} t^{-1/2}) + O(\varsigma(1-\gamma)^{-2})$, one factor $(1-\gamma)^{-1}$ beyond the ergodic average and the source of $(1-\gamma)^{-6}$ in K.

Certification and convergence are proved together without conditioning on a future good event. A virtual process that switches to exact Q-values at the first certification failure satisfies the bias conditions unconditionally. Its Bregman potential stays below the level that forces $\pi_t(a^*(s)|s) \ge \underline{\pi}$ (Lemma G.8), so it never switches and coincides with the algorithm on a high-probability event.

5.2 Coordinate locking and phase accounting

The analysis of Ju & Lan (2025) guarantees $\pi_t(a^*(s)|s) \ge \underline{\pi}$ and treats all states identically, but states with large gaps concentrate on the optimal action much faster. The route through the value error (Xiao, 2022; Li et al., 2024) bounds the mass $1 - \pi_t(a^*(s)|s)$ by $\varepsilon_t/\Delta(s)$ (Lemma D.1); at an active state this mass is only known to exceed $\underline{\pi}$, so that route pays $\underline{\pi}^{-1}$ at every accuracy level. The cumulative form avoids the division: relative to $a^*(s)$, the dual coordinate of an action is minus its accumulated estimated advantage, and it cannot fall below a fixed level while the action is non-rare.

Lemma 5.3 (Coordinate locking)**.** *On the event $\mathcal{E}_{\mathrm{opt}}$ of Lemma G.3, which has probability at least $1 - \delta/8$, for all $1 \le t \le K$, all s and every action with $\pi_t(a|s) \ge \underline{\pi}$,*

$$Q^{\pi^*}(s,a) - V^{\pi^*}(s) \le u_t := \min\Big\{ (1-\gamma)^{-1},\ \frac{4(1-\gamma)A_K}{\sqrt{t}} + 2(1-\gamma)\varepsilon \Big\},$$

where A_K is the constant of Appendix G. In particular $\mathcal{S}_t^{\mathrm{multi}} \subseteq \{s : \Delta(s) \le u_t\}$.

The relative bias of Lemma 5.2 holds for every action, including actions that were rare earlier, so the proof needs no history of an action's status. Since $(1-\gamma)A_K = \widetilde{O}(|\mathcal{A}|^{(1-p)/2}(1-\gamma)^{-2})$, the locking scale decays like the one-step advantage and carries no factor $\underline{\pi}^{-1}$.

Part B then bounds each rollout. Certification places every iterate in Π_θ, so each non-rare pair (s,a) is hit within $L_\gamma(\varsigma) + \lceil 2\theta^{-1} H_\theta(s) \ln(e/\delta') \rceil$ steps with probability $1 - \delta'$ (Lemma C.1), without the mixing time or the stationary distribution of π_t. Lemma 5.3 restricts the active set to $\{s : \Delta(s) \le u_t\}$, and summing over the accuracy phases of u_t yields $\mathcal{I}_{\mathrm{ph}}(H_\theta; (1-\gamma)\varepsilon)$, with one factor θ^{-1} from pair hitting and none from locking.

6 Discussion

The exploration cost of single-trajectory SPMD is controlled by the certified hitting cost H_θ: charging H_θ only at unresolved states gives an exploration term governed by $\mathcal{I}_\infty(H_\theta; (1-\gamma)\varepsilon)$ for every MDP with an irreducible deterministic optimal policy. On an action-dependent family, a lower bound with deterministic costs matches the upper bound up to logarithmic factors for fixed γ. Three questions remain open. First, H_θ can be exponential in the length ℓ of unavoidable paths through action-dependent states, and we do not know when the factor $\theta^{-\ell}$ can be removed from the upper bound. Second, the iteration count carries $(1-\gamma)^{-6}$, two powers more than the last-iterate analysis of Ju & Lan (2025) would give, because that analysis needs bias control along earlier iterates (Remark 3.8). Whether the two powers can be recovered is open. Third, the lower bound is family-specific and, like $\mathcal{I}_\infty$, takes a maximum over states; general lower bounds in terms of hitting costs and gaps remain open, as do lower bounds that sum over states. The analysis is tabular with $h = 0$.

AI USE STATEMENT

The human authors led the research: they chose the problem, identified the lines of work and the analytical tools the paper builds on, and fixed the formulation. They reached the formulation iteratively, having AI models conjecture the resulting bounds and test them in simulations, and revising it after each mismatch. The authors directed the derivations, carried out mainly with GPT (OpenAI), turned the model outputs into the arguments presented here, and verified the results. They set the outline and the line of argument; under their direction, Claude (Anthropic) wrote drafts, which they corrected, and set up and ran the numerical illustration of Appendix A. No data sets were generated with AI tools, no human-subject data are involved, and the authors take full responsibility for all content.

REPRODUCIBILITY STATEMENT

Algorithm 1 and all its parameters are specified in Section 3.1. Assumption 2.1 (Section 2) is the only structural assumption of the upper bounds, and Assumption 3.10 (Section 3.5) is used only for Corollary 3.11. Complete proofs of all results are in the appendix, which opens with a guide to its structure. Appendix A lists the parameters, seeds and budgets of the numerical illustration; the code, which reproduces Figure 2, will be released upon publication.

REFERENCES

Alekh Agarwal, Sham Kakade, and Lin F Yang. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020.

Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in Markov decision processes. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Conference on Learning Theory*, pp. 41–53, 2010.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2):235–256, 2002.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3): 325–349, 2013.

Prateek Jain, Dheeraj M Nagaraj, and Praneeth Netrapalli. Making the last iterate of SGD information theoretically optimal. *SIAM Journal on Optimization*, 31(2):1108–1130, 2021.

Caleb Ju and Guanghui Lan. Auto-exploration for online reinforcement learning. *arXiv preprint arXiv:2512.06244 (version 4)*, 2025.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical Programming*, 198(1):1059–1106, 2023.

Yan Li and Guanghui Lan. Policy mirror descent inherently explores action space. *SIAM Journal on Optimization*, 35(1):116–156, 2025.

Yan Li, Guanghui Lan, and Tuo Zhao. Homotopic policy mirror descent: policy convergence, algorithmic regularization, and improved sample complexity. *Mathematical Programming*, 207(1): 457–513, 2024.

Phalguni Nanda and Zaiwei Chen. A minimal-assumption analysis of Q-learning with time-varying policies. *arXiv preprint arXiv:2510.16132*, 2025.

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

Andrea Tirinzoni, Aymen Al Marjani, and Emilie Kaufmann. Near instance-optimal PAC reinforcement learning for deterministic MDPs. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Andrea Tirinzoni, Aymen Al Marjani, and Emilie Kaufmann. Optimistic PAC reinforcement learning: the instance-dependent view. In *International Conference on Algorithmic Learning Theory*, pp. 1460–1480. PMLR, 2023.

Andrew Wagenmaker and Kevin Jamieson. Instance-dependent near-optimal policy identification in linear MDPs via online experiment design. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Andrew Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent PAC reinforcement learning. In *Conference on Learning Theory*, pp. 358–418. PMLR, 2022.

Lin Xiao. On the convergence rates of policy gradient methods. *Journal of Machine Learning Research*, 23(282):1–36, 2022.

APPENDIX

How to read this appendix. Appendices A and B are auxiliary and are not needed for the proofs. Appendix C proves the results of Sections 3.2 and 3.5 and the certified hitting lemma used in Part B. The next appendices prove the lemmas of Section 5, Part A of Theorem 3.7 (convergence and certification, Appendix G) and Part B (sample complexity, Appendix H). The remaining appendices contain the lower bounds, the action-dependent star, the results on the ring, the proof of Proposition 3.5, and additional related work. All SPMD conditions and hitting-time lemmas used in the proofs are restated here, so the paper can be read independently. Throughout, $\kappa_s^\pi(q) := (1-\gamma)\sum_{t=0}^{\infty}\gamma^t P_\pi^{(t)}(s,q)$ denotes the discounted state-occupancy measure of π started at s.

Conventions. Throughout, all logarithms are base 2 unless explicitly stated otherwise. We write $a \lesssim b$ to mean $a \le Cb$ for a universal numerical constant $C > 0$ (similarly $\gtrsim$).

Notation. $\Lambda_{\mathrm{hp}} := O(\log^2(K|\mathcal{S}||\mathcal{A}|/(\delta(1-\gamma)\underline{\pi})))$ denotes a universal polylogarithmic factor that absorbs high-probability union bounds. We write $\theta := \underline{\pi}$ and use the shorthand $\mathcal{S}_t^{\mathrm{multi}}$ (observable active set: states where π_t assigns $\ge \underline{\pi}$ to ≥ 2 actions); $\{s : \Delta(s) \le u_t\}$ (analytic unresolved set, Lemma 5.3); $\mathrm{WB}_t(s) := \mathbb{E}[\langle -\xi_t(s,\cdot), \pi_t(\cdot|s) - \pi^*(\cdot|s)\rangle \mid \mathcal{F}_t]$ (weighted bias).

A NUMERICAL ILLUSTRATION

Setup. The experiment tests the mechanism behind the separation of Section 4 on the algorithm itself rather than on an idealized charge: once the bottleneck locks, the local stopping rule of Algorithm 1 stops paying for it, while the global rule of Ju & Lan (2025) keeps paying about $1/r$ per iteration. It uses TOMC rollouts on one continuous trajectory starting at g_2, deterministic costs, the Tsallis-p mirror step with $p = 1/\log_2((1-\gamma)^{-1})$, stepsizes $\eta_t = 2/\sqrt{t+1}$, rare-action threshold $\underline{\pi} = 0.05$, and residual threshold $\varsigma(1-\gamma) = 10^{-4}$, so that $L_\gamma(\varsigma) = 88$. The two stopping rules share everything else, and exact values are computed only for evaluation. The gaps are $\Delta(b) = 0.35$ and $\Delta(g_i) = 0.09$. The control lowers the probability that R stays at b from $\frac{1}{2}$ to 0.1, which gives $\Delta(b) = 0.07 < \Delta(g_i)$. These are practical parameter choices for illustrating the mechanism, not the constants of Theorem 3.7. The seeds are $0, \dots, 4$. Each run stops at a fixed interaction budget, $2500(1.3/r + 900)$ for the family (i.e., $55{,}498{,}000$ at $1/r = 16384$) and 1.5 times that for the control. Panel (c) reports, for each accuracy level, the median over seeds of the interactions needed to reach it, counting a seed that never reaches the level as $+\infty$. The script `experiment_ring.py` reproduces the figure in about 16 minutes on a laptop CPU.

What the figure shows. Once the bottleneck locks (between iterations 6 and 116 across seeds), the local rollout falls to about 720 steps, while the global rollout stays at about $11{,}300$ steps, of order

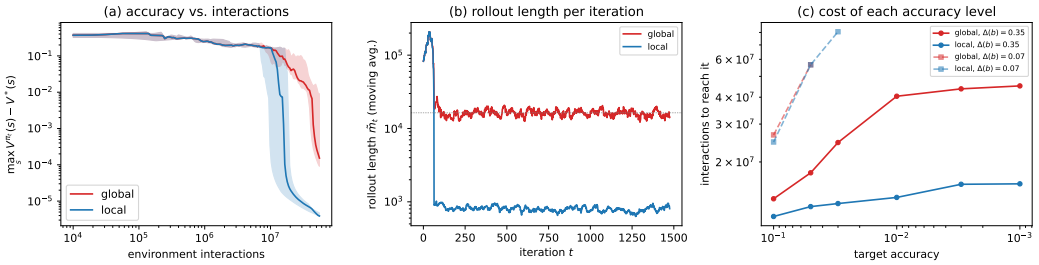


Figure 2: Algorithm 1 on one continuous trajectory of the ring of Example 4.1, with the jump probability r set independently of n ($n = 8$, $r = 1/16384$, $\eta = 0.05$, $\gamma = 0.9$); medians and interquartile ranges over five seeds, each run stopped at a common interaction budget. **(a)** Value error against environment interactions. **(b)** Rollout length per iteration (moving average over 25 iterations): once the bottleneck locks, the local rollout falls to the cost of covering the ring, while the global rollout keeps paying about $1/r$ (dotted) per iteration. **(c)** Interactions needed to reach each accuracy level, for the family ($\Delta(b) = 0.35$) and for a control in which the bottleneck is hard to resolve ($\Delta(b) = 0.07$, faded). A control point is shown when at least three seeds reach that level within the budget.

$1/r$. The end-to-end saving is smaller than the per-iteration one. Before b locks, both rules must hit (b, R) while its probability falls toward $\underline{\pi}$, at a cost of order $1/(r\underline{\pi})$ per iteration and about 11 million interactions in total here. This is the factor $\underline{\pi}^{-1}$ of Lemma C.1. After that, each further accuracy level costs the local rule much less: going from error 0.1 to 0.001 takes 4.7 million more interactions for the local rule and 32 million for the global one (medians), so the ratio of total interactions grows from 1.2 at error 0.1 to 2.9 at error 0.001. In the control, the bottleneck does not resolve before the generic states, the two rules use rollouts of the same length, and their costs coincide.

Why vary $1/r$ rather than n. At $r = n^{-2}$ and $n = 8$ the two rules coincide up to noise, because covering the ring costs more than $1/r = 64$ steps per iteration; this cost includes actions whose probability is only slightly above $\underline{\pi}$. The separation in n requires $n^2 \gg n \log n/\underline{\pi}$, beyond the reach of a small simulation. Varying $1/r$ at fixed n isolates the part of the cost that the local rule removes: in a sweep over $1/r \in \{64, 256, 1024, 4096, 16384\}$ (`-sweep`), the post-lock local rollout stayed between 560 and 870 steps, while the global one grew from 830 to 11,300. The theory covers this extension: the proof of Lemma 4.2 uses $r = n^{-2}$ only through $r \le n^{-2}$. For $0 < r \le n^{-2}$ it gives $H_\theta(b) = 1/r$ and $H_\theta(s) \le 64n/3$ for $s \ne b$, hence an exploration term $\widetilde{O}(r^{-1} + n\varepsilon^{-2})$ for fixed γ.

Bias tolerance. In a pilot run with $\varsigma(1-\gamma) = 10^{-2}$, i.e., $\varsigma = 0.1 \approx \Delta(g_i)$, the allowed bias was of the order of the gap. The local rule then converged more slowly per iteration and lost its advantage. This is the regime excluded by the requirement $\varsigma = \Theta((1-\gamma)^2\varepsilon)$.

B On-the-fly implementation and a shrinking active set

On-the-fly algorithm (the on-the-fly variant of Algorithm 1). Maintain a label $\ell(s) \in \{\texttt{unknown}, \texttt{single}, \texttt{multi}\}$ for each state (initially all `unknown`) and a set U of non-rare pairs not yet covered. At each rollout step m:

(i) When the current state s_m has $\ell(s_m) = \texttt{unknown}$, scan $\pi_t(\cdot|s_m)$ once. Set

$$\ell(s_m) \leftarrow \begin{cases} \texttt{multi} & \text{if } \left|\{a : \pi_t(a|s_m) \ge \underline{\pi}\}\right| \ge 2 \quad \text{(add its non-rare pairs to } U\text{)}, \\ \texttt{single} & \text{otherwise.} \end{cases}$$

(ii) Remove from U any (s, a) satisfying

$$h_m^{\pi_t}(s, a) \le \varsigma(1-\gamma).$$

(iii) When $U = \varnothing$, perform a *zero-interaction sweep*: scan all remaining unknown policy rows, classifying each as single or multi. If new multi-action pairs are added to U, continue the rollout; otherwise stop.

Proposition B.1 (Tabular on-the-fly implementation)**.** *Fix iteration t and policy π_t. Consider the on-the-fly subroutine in Algorithm 1, which (i) classifies each state s by scanning $\pi_t(\cdot|s)$ at most once (either upon its first visit in the rollout or during the zero-interaction sweep), (ii) stores first-visit times for all visited pairs, and (iii) stops when all states have been classified and all non-rare pairs at multi-action states satisfy the bias threshold. Then the returned stopping time equals the definition in* (3)*: $\hat{m}_t = \tilde{m}_t^{\mathrm{loc}}$. The additional tabular bookkeeping is at most $O(|\mathcal{S}||\mathcal{A}|)$ per SPMD iteration.*

Proof. The state classification depends only on the fixed policy table π_t: once all states have been classified, $\ell(s) = \texttt{multi}$ if and only if $s \in \mathcal{S}_t^{\mathrm{multi}}$. The zero-interaction sweep ensures that all states are classified no later than the first time U becomes empty: at that point, any remaining unknown states are scanned without consuming environment interactions. If the sweep discovers new multi-action states, their non-rare pairs are added to U and the rollout continues; otherwise, all multi-action pairs have been covered and the subroutine stops. Since first-visit times $\tau_m^{\pi_t}(s,a)$ are recorded for all visited pairs regardless of classification order, the set U_m of pairs not yet covered at time m satisfies

$$U_m = \varnothing \quad \Longleftrightarrow \quad \max_{\substack{(s,a)\in\mathcal{Z}_{\mathrm{nr}}^{\pi_t}(\underline{\pi}) \\ s\in\mathcal{S}_t^{\mathrm{multi}}}} h_m^{\pi_t}(s,a) \le \varsigma(1-\gamma).$$

Therefore $\hat{m}_t = \tilde{m}_t^{\mathrm{loc}}$. The bookkeeping cost is at most $|\mathcal{S}|$ policy-row scans of $O(|\mathcal{A}|)$ each. □

The multi-action set can only shrink across iterations.

Lemma B.2 (Rare actions stay rare)**.** *For every t, s and a: if $\pi_t(a|s) < \underline{\pi}$, then $\pi_{t+1}(a|s) \le \pi_t(a|s)$. Consequently the set of non-rare actions at each state is non-increasing in t, and so is $\mathcal{S}_t^{\mathrm{multi}}$: a state that leaves $\mathcal{S}_t^{\mathrm{multi}}$ never returns.*

Proof. Write $\bar{Q} := (1-\gamma)^{-1}$ and $\omega'(x) := -x^{p-1}/(1-p)$, which is strictly increasing on $(0,1]$ and tends to $-\infty$ as $x \to 0$. The mirror step at s therefore has an interior solution, characterized by $\omega'(\pi_{t+1}(b|s)) = \omega'(\pi_t(b|s)) - \eta_t \tilde{Q}^{\pi_t}(s,b) + \lambda_t$ for all $b \in \mathcal{A}$ and a scalar λ_t. Every estimate lies in $[0,\bar{Q}]$, and every rare action receives exactly $\bar{Q}$ by (1). If $\lambda_t > \eta_t \bar{Q}$, every coordinate of $\pi_{t+1}(\cdot|s)$ would exceed the corresponding coordinate of $\pi_t(\cdot|s)$, which is impossible because both rows sum to 1. Hence $\lambda_t \le \eta_t \bar{Q}$, and a rare action a satisfies $\omega'(\pi_{t+1}(a|s)) - \omega'(\pi_t(a|s)) = \lambda_t - \eta_t \bar{Q} \le 0$. □

The argument is pathwise and uses only $h = 0$, the fixed threshold $\underline{\pi}$ and the common value $\bar{Q}$ of the rare-action estimates. In particular, a state classified `single` at some iteration stays single at all later iterations, so its row need not be rescanned.

C Certified hitting (Sections 3.2 and 3.5)

Proof of Lemma 3.9. Fix $\pi \in \Pi_\theta$ and x, s. If $x = s$, the first two quantities equal 1 and the claim holds. Otherwise, for $t \le \ell$ the event $\{\tau_s = t\}$ is the disjoint union, over paths $x_0 = x, x_1, \ldots, x_t = s$ with $x_k \ne s$ for $k < t$, of the events $\{s_0 = x_0, \ldots, s_t = x_t\}$. Along such a path, if $x_k \notin \mathcal{S}_{\mathrm{dep}}$ then $P_\pi(x_k, x_{k+1}) = P(x_{k+1}|x_k) = P_{\pi^*}(x_k, x_{k+1})$; if $x_k \in \mathcal{S}_{\mathrm{dep}}$ then $P_\pi(x_k, x_{k+1}) \ge \pi(a^*(x_k)|x_k)\, P(x_{k+1}|x_k, a^*(x_k)) \ge \theta P_{\pi^*}(x_k, x_{k+1})$. Multiplying along the path,

$$\Pr_x^{\pi}(s_1 = x_1, \ldots, s_t = x_t) \ge \theta^{|\{k<t:\, x_k \in \mathcal{S}_{\mathrm{dep}}\}|} \Pr_x^{\pi^*}(s_1 = x_1, \ldots, s_t = x_t).$$

Summing over first-passage paths of length $t \le \ell$ gives the first inequality; the second follows from $N_{\tau_s} \le \tau_s \le \ell$ on $\{\tau_s \le \ell\}$. □

Proof of Proposition 3.2. Lower bound. Fix $\ell \ge 1$ with $q := q_s^\theta(\ell) > 0$; if there is no such ℓ, then $H_\theta(s) = \infty$ and there is nothing to prove. For $\pi \in \Pi_\theta$ and any x, the Markov property at times $\ell, 2\ell, \ldots$ and the uniformity of q over starting states give $\Pr_x^\pi(\tau_s > k\ell) \le (1-q)^k$. Hence $\mathbb{E}_x^\pi[\tau_s] = \sum_{t\ge 0} \Pr_x^\pi(\tau_s > t) \le \ell \sum_{k\ge 0}(1-q)^k = \ell/q$. Taking the supremum over π and x and the infimum over ℓ gives $D_\theta(s) \le H_\theta(s)$.

Upper bound. Assume $D_\theta(s) < \infty$ and let $\ell := \max\{1, \lceil 2D_\theta(s) \rceil\}$. By Markov's inequality, $\Pr_x^\pi(\tau_s > \ell) \le D_\theta(s)/\ell \le \frac{1}{2}$ for every $\pi \in \Pi_\theta$ and every x, so $q_s^\theta(\ell) \ge \frac{1}{2}$ and $H_\theta(s) \le 2\ell \le 4D_\theta(s) + 2$.

Finiteness. If P_{π^*} is irreducible, then from every x there is a path of positive P_{π^*}-probability to s of length at most $|\mathcal{S}|$, so $\Pr_x^{\pi^*}(\tau_s \le |\mathcal{S}|) \ge p_{\min} > 0$ uniformly in x, where $p_{\min}$ is the smallest of these finitely many path probabilities. Lemma 3.9 then gives $q_s^\theta(|\mathcal{S}|) \ge \theta^{|\mathcal{S}|} p_{\min} > 0$. Conversely, $\pi^* \in \Pi_\theta$, so $H_\theta(s) < \infty$ for all s forces $\Pr_x^{\pi^*}(\tau_s < \infty) > 0$ for all x, s, which is irreducibility. □

Part B needs a high-probability rollout length for a single pair. For a stationary policy π, a pair (s, a) and $\delta' \in (0, 1)$, let $\mathrm{Hit}_{\pi,\delta',\varsigma}(s, a)$ be the smallest rollout length that drives the discounted hitting residual $h_m^\pi(s, a)$ below $\varsigma(1-\gamma)$ with probability at least $1-\delta'$ from every starting state; Appendix H states this definition formally.

Lemma C.1 (Certified pair hitting)**.** *Let $\pi \in \Pi_\theta$, let (s, a) satisfy $\pi(a|s) \ge \theta$, and let $\delta' \in (0, 1)$ and $\varsigma(1-\gamma) \in (0, 1)$. Then, with L_γ as in Theorem 3.7,*

$$\mathrm{Hit}_{\pi,\delta',\varsigma}(s, a) \le L_\gamma(\varsigma) + \left\lceil \frac{2H_\theta(s)}{\theta} \ln \frac{e}{\delta'} \right\rceil.$$

Proof. Fix $\ell \ge 1$ with $q := q_s^\theta(\ell) > 0$, and set $B := \ell + 1$ and $T_r := rB$ for $r \ge 0$. Let $\mathcal{G}_r$ be the σ-field generated by the trajectory up to the state s_{T_r}, before the action at time T_r is drawn, and let E_r be the event that the pair (s, a) is visited at some time in $\{T_r, \dots, T_r + \ell\}$. By the Markov property, the first visit to s after T_r occurs within ℓ steps with conditional probability at least $\Pr_{s_{T_r}}^\pi(\tau_s \le \ell) \ge q$, and at that visit the action a is drawn with probability $\pi(a|s) \ge \theta$; hence $\Pr(E_r \mid \mathcal{G}_r) \ge \theta q$. Since $E_0, \dots, E_{r-1}$ are $\mathcal{G}_r$-measurable, induction gives $\Pr(E_0^c \cap \cdots \cap E_{k-1}^c) \le (1 - \theta q)^k \le \delta'$ for $k := \lceil \ln(1/\delta')/(\theta q) \rceil$. Off this event the first visit τ to (s, a) satisfies $\tau < kB$, so with $M := kB + L_\gamma(\varsigma)$ we get $h_M^\pi(s, a) = \gamma^{M-\tau} \le \gamma^{L_\gamma(\varsigma)} \le \varsigma(1-\gamma)$. Thus $\mathrm{Hit}_{\pi,\delta',\varsigma}(s, a) \le M$. Finally, $B \le 2\ell$ and $\theta q \le 1$ give $kB \le (\ln(1/\delta')/(\theta q) + 1) \cdot 2\ell \le \frac{2\ell}{\theta q} \ln \frac{e}{\delta'}$. The infimum of $\ell/q_s^\theta(\ell)$ over $\ell \ge 1$ is attained because $\ell/q_s^\theta(\ell) \ge \ell$; choosing a minimizing ℓ gives the claim. □

Proof of Corollary 3.11. By Assumption 3.10 and the choice of $\bar{b}$, $|P_{\pi^*}^{(\bar{b})}(x, y) - \nu^*(y)| \le C^*(\rho^*)^{\bar{b}} \le \underline{\nu}^*/2 \le \nu^*(y)/2$, so $\Pr_x^{\pi^*}(\tau_s \le \bar{b}) \ge P_{\pi^*}^{(\bar{b})}(x, s) \ge \nu^*(s)/2$ for all x. Lemma 3.9 with $\ell = \bar{b}$ gives $q_s^\theta(\bar{b}) \ge \theta^{\bar{b}} \nu^*(s)/2$, hence $H_\theta(s) \le 2\bar{b}/(\theta^{\bar{b}} \nu^*(s))$. Substituting $H_\theta \le 2\bar{b}\theta^{-\bar{b}} \cdot (1/\nu^*)$ into Theorem 3.7 and using Proposition 3.5 gives the second claim. □

Remark C.2 (Comparison with Ju & Lan (2025), factor by factor)**.** We compare explicit factors with the bound as stated in Theorem 3.5 of the arXiv version 4 of Ju & Lan (2025), specialized to $h = 0$. In the worst case, when all gaps are at most $(1-\gamma)\varepsilon$, Theorem 3.7 and Corollary 3.11 give the exploration term $\widetilde{O}(|\mathcal{A}|^{1-p}\, \bar{b}\, \theta^{-(\bar{b}+1)}(1-\gamma)^{-6}\varepsilon^{-2}/\underline{\nu}^*)$, with $\theta = \underline{\pi}$. Their bound is $\widetilde{O}(|\mathcal{A}|^{2-p_5}(1-\gamma)^{-5}\varepsilon^{-2}L_5^2\, D_{\mathrm{expl}})$, where $L_5 := \log(1/\delta)$, $p_5 := 1/(2 + \log(1/(1-\gamma)))$ and $\bar{b}_5 := \lceil \log(8/\underline{\nu}^*)/\log(1/\rho^*) \rceil$. In terms of their threshold $\theta_5 := (1-\gamma)p_5^2/(100|\mathcal{A}|L_{k,5})$, where $L_{k,5} := \log(2|\mathcal{S}|k/\delta)$ and k is their iteration count,

$$D_{\mathrm{expl}} = \Big(\frac{|\mathcal{A}|L_5}{1-\gamma}\Big)^{2\bar{b}_5} \frac{1}{(1-\rho^*)(\underline{\nu}^*)^3} = \theta_5^{-2\bar{b}_5} \Big(\frac{p_5^2 L_5}{100 L_{k,5}}\Big)^{2\bar{b}_5} \frac{1}{(1-\rho^*)(\underline{\nu}^*)^3}.$$

Their value guarantee is $\varepsilon \log(4k) \log(2|\mathcal{S}| \log(4k)/\delta)$ rather than ε. Reparameterizing the accuracy changes the stated bound only by polylogarithmic factors, because D_{expl} does not depend on k (the factors $L_{k,5}$ in the second form cancel). The explicit factors differ in three ways:

- the inverse threshold enters as $\theta^{-(\bar{b}+1)}$ here, and as $\theta_5^{-2\bar{b}_5}$ times $(p_5^2 L_5/(100 L_{k,5}))^{2\bar{b}_5}$ there;
- outside the exponents, the minimum stationary probability enters as $(\underline{\nu}^*)^{-1}$ here and as $(\underline{\nu}^*)^{-3}$ there; it also enters both bounds through $\bar{b}$ and $\bar{b}_5$;
- the mixing rate appears here only through $\bar{b}$, and there through $\bar{b}_5$ and the explicit factor $(1-\rho^*)^{-1}$.

The thresholds and entropic indices also differ, and the second factor in the first item is exponential in $\bar{b}_5$. These displayed powers alone do not establish a uniform ordering between the two bounds.

Proof of Corollary 3.12. If no kernel depends on the action, every stationary policy induces the same state process, so $\Pr_x^\pi(\tau_s \le \ell) = \Pr_x(\tau_s \le \ell)$ for all π, and $q_s^\theta(\ell) = \inf_x \Pr_x(\tau_s \le \ell)$. □

D STATEWISE LOCKING THROUGH THE VALUE ERROR

The following lemma is the value-based route mentioned in Section 5.2. Part B does not use it; the lower bounds of Appendix I use it to turn ε-optimality into a test.

Lemma D.1 (Statewise locking)**.** *Let π be a stationary policy and s a state with $\Delta(s) > 0$. If $V^\pi(s) - V^{\pi^*}(s) \le \varepsilon'$, then $1 - \pi(a^*(s)|s) \le \varepsilon'/\Delta(s)$.*

Proof. By the performance difference lemma (Ju & Lan, 2025, Lemma 2.1),

$$V^\pi(s) - V^{\pi^*}(s) = \frac{1}{1-\gamma} \sum_{q \in \mathcal{S}} \kappa_s^\pi(q)\, \psi^{\pi^*}(q, \pi(\cdot|q)),$$

where $\psi^{\pi^*}(q, v) = \langle Q^{\pi^*}(q, \cdot), v\rangle - V^{\pi^*}(q) \ge 0$ for all q and v, by optimality of π^*. Every term is non-negative and $\kappa_s^\pi(s) \ge 1 - \gamma$ (the $t = 0$ term), so $V^\pi(s) - V^{\pi^*}(s) \ge \psi^{\pi^*}(s, \pi(\cdot|s))$. Moreover

$$\psi^{\pi^*}(s, \pi(\cdot|s)) = \sum_{a \ne a^*(s)} \pi(a|s)\big[Q^{\pi^*}(s, a) - V^{\pi^*}(s)\big] \ge \Delta(s)\big(1 - \pi(a^*(s)|s)\big),$$

and combining with $V^\pi(s) - V^{\pi^*}(s) \le \varepsilon'$ gives the claim. □

E PROOF OF LEMMA 5.1 (ONE-SIDED BIAS) AND REWARD MAXIMIZATION

Proof. Fix t and condition on $\mathcal{F}_t$. Extend the rollout of iteration t indefinitely under the fixed policy π_t. Since the Tsallis iterates are interior, $P_{\pi_t}(x, y) \ge \pi_t(a^*(x)|x)P(y|x, a^*(x)) > 0$ whenever $P_{\pi^*}(x, y) > 0$, so P_{π_t} is irreducible under Assumption 2.1 and the untruncated first visit $\tau_z := \min\{u \ge 0 : Z_u = z\}$ is finite almost surely. Let $G_z := \sum_{u \ge 0} \gamma^u C_{\tau_z + u}$ be the full realized return from that visit; it is bounded by $(1 - \gamma)^{-1}$. By the tower property and $\mathbb{E}[C_u \mid \mathcal{H}_u] = c(Z_u)$, $\mathbb{E}[G_z \mid \mathcal{H}_{\tau_z}] = \sum_{u \ge 0} \gamma^u\, \mathbb{E}[c(Z_{\tau_z + u}) \mid \mathcal{H}_{\tau_z}]$. Given $(\mathcal{H}_u, C_u)$, the next state depends on the past only through Z_u, and actions are drawn from $\pi_t(\cdot|S_u)$; so after the stopping time τ_z the state-action process is the chain of π_t started at z, and the right-hand side equals $Q^{\pi_t}(z)$. Since $\mathcal{F}_t \subseteq \mathcal{H}_{\tau_z}$, also $\mathbb{E}[G_z \mid \mathcal{F}_t] = Q^{\pi_t}(z)$. If z is visited before the stopping time m, the TOMC estimate is the truncation of G_z at m; otherwise it is 0. In both cases $0 \le \tilde{Q}^{\pi_t}(z) \le G_z$ pathwise, because observed costs are non-negative, and taking conditional expectations gives the claim. □

Remark E.1 (Reward maximization)**.** With observed rewards $R_t \in [0, 1]$, run the algorithm on $C_t := 1 - R_t$. Then $Q_C^\pi = (1 - \gamma)^{-1} - Q_R^\pi$ and $V_C^\pi = (1 - \gamma)^{-1} - V_R^\pi$ for every π, so the optimal policies and the gaps $\Delta(s)$ coincide, and every result applies verbatim. In reward coordinates the same estimator reads $\tilde{Q}_R(z) = \sum_{t=\tau_m(z)}^{m-1} \gamma^{t-\tau_m(z)} R_t + \gamma^{m-\tau_m(z)}/(1 - \gamma)$ for non-rare z and $\tilde{Q}_R(z) = 0$ for rare z, because $\tilde{Q}_R = (1-\gamma)^{-1} - \tilde{Q}_C$ pathwise. The second term fills the unobserved tail optimistically. Plain truncation of rewards, without it, would reverse the sign used in Lemma 5.1.

F PROOF OF LEMMA 5.2 (RELATIVE BIAS)

The proof converts the residual condition of the stopping rule into a bias bound.

Lemma F.1 (Residual-to-bias conversion)**.** *Fix iteration t and condition on $\mathcal{F}_t$. If (s, a) is non-rare and the localized stopping rule certifies $h_{m_t}^{\pi_t}(s, a) \le \varsigma(1 - \gamma)$, then $|\mathbb{E}[\xi_t(s, a) \mid \mathcal{F}_t]| \le \varsigma$.*

Proof. If $\varsigma(1-\gamma) \ge 1$ the claim is trivial: $|\xi_t(s, a)| \le \bar{Q} = (1-\gamma)^{-1} \le \varsigma$. Otherwise $\varsigma(1-\gamma) < 1$, and since $h_m^{\pi_t}(s, a) = 1$ when the pair is not hit, the residual condition $h_{m_t}^{\pi_t}(s, a) \le \varsigma(1 - \gamma) < 1$ implies $\tau_m^{\pi_t}(s, a) < m_t$ a.s. With the untruncated first visit and the full return G_z of the proof of Lemma 5.1, the conditional bias is the expected missing tail, and the tail is bounded pathwise by the residual, $G_z - \tilde{Q}_{m_t, \underline{\pi}}(s, a) = \sum_{r \ge m_t} \gamma^{r - \tau_m} C_r \le h_{m_t}^{\pi_t}(s, a)/(1 - \gamma)$. Hence $0 \le Q^{\pi_t}(s, a) -$

$\mathbb{E}[\tilde{Q}_{m_t,\underline{\pi}}(s,a) \mid \mathcal{F}_t] = \mathbb{E}[G_z - \tilde{Q}_{m_t,\underline{\pi}}(s,a) \mid \mathcal{F}_t] \le (1-\gamma)^{-1}\,\mathbb{E}[h^{\pi_t}_{m_t}(s,a) \mid \mathcal{F}_t] \le \varsigma$, where the last step uses $h^{\pi_t}_{m_t}(s,a) \le \varsigma(1-\gamma)$ almost surely. □

Proof. All expectations are conditional on $\mathcal{F}_t$. Write $a^* = a^*(s)$ and recall $b_t(s,a) = \mathbb{E}[\xi_t(s,a)]$ with $\xi_t = \tilde{Q}^{\pi_t} - Q^{\pi_t}$. Three facts hold at every state s: (a) if (s,a) is non-rare, then $b_t(s,a) \le 0$ by Lemma 5.1; (b) if (s,a) is non-rare and $s \in \mathcal{S}^{\mathrm{multi}}_t$, then $b_t(s,a) \ge -\varsigma$ by the stopping rule (3) and Lemma F.1; (c) if (s,a) is rare, then $\tilde{Q}^{\pi_t}(s,a) = (1-\gamma)^{-1} \ge Q^{\pi_t}(s,a)$, so $b_t(s,a) \ge 0$. By hypothesis every rare action differs from a^*, so (s,a^*) is non-rare and $b_t(s,a^*) \le 0$ by (a). Fix $a \ne a^*$. If a is rare, then $b_t(s,a) - b_t(s,a^*) \ge 0$ by (c) and (a). If a is non-rare, then s has two non-rare actions, so $s \in \mathcal{S}^{\mathrm{multi}}_t$, and $b_t(s,a) - b_t(s,a^*) \ge -\varsigma$ by (b) and (a). This proves $b_t(s,a) - b_t(s,a^*) \ge -\varsigma$ for all a, with lower bound 0 when $s \notin \mathcal{S}^{\mathrm{multi}}_t$.

For $v \in \Delta_{|\mathcal{A}|}$, since $\pi^*(\cdot|s)$ is the point mass at a^*,

$$\langle b_t(s,\cdot), v - \pi^*(\cdot|s)\rangle = \sum_a v(a)\big(b_t(s,a) - b_t(s,a^*)\big) \ge -\varsigma,$$

with lower bound 0 at single-action states. Taking $v = \pi_t(\cdot|s)$ gives the weighted-bias bounds. □

G Convergence proof (Part A of Theorem 3.7) and coordinate locking

Relation to the last-iterate analysis of Ju & Lan (2025). The high-probability last-iterate bound of Ju & Lan (2025) is derived from a weighted-deviation lemma (Lemma A.1 of its arXiv version 4, used in Lemma A.2 there). Its proof bounds the conditional bias along $\pi_t - \pi_{t_0}$ for earlier iterates π_{t_0}, whereas the stated assumption controls the direction $\pi_t - \pi^*$. The two coincide when the bias vector is small in every coordinate, which the localized estimator does not provide: a suboptimal action that has become rare carries the pessimistic default and may have more mass under π_{t_0}. We therefore do not use these lemmas. The proof below compares only with the deterministic π^*, which Lemma 5.2 controls for every policy row (Section 5.1).

Plan. The proof runs on a virtual process that switches to exact Q-values at the first certification failure (Lemma G.1). Summing its mirror steps gives the cumulative form (Lemma G.2), and one event $\mathcal{E}_{\mathrm{opt}}$ controls the noise (Lemma G.3). On this event, the one-step inequality of Lemma G.4 yields an ergodic value bound and a Bregman potential (Lemma G.5). The cumulative form turns the ergodic bound into a last-iterate bound (Lemma G.6), and the potential certifies every iterate (Lemmas G.7 and G.8). Proposition G.9 combines these. The section ends with the proof of Lemma 5.3, which reuses the event and the ergodic bound.

Notation. Write $\bar{Q} := (1-\gamma)^{-1}$ and $\theta := \underline{\pi}$, and let $D_0 := (|\mathcal{A}|^{1-p} - 1)/(p(1-p))$, which equals $D^{\pi^*}_{\pi_0}(q)$ at every state q because π_0 is uniform and π^* is deterministic. For the stepsizes $\eta_u = \alpha/\sqrt{u+1}$ let $W_t := \sum_{u<t}\eta_u$ and $S_t := \sum_{u<t}\eta_u^2$; for $t \ge 1$, $\alpha\sqrt{t} \le W_t \le 2\alpha\sqrt{t}$ and $S_t \le \alpha^2\ln(et)$. For a sequence of policies (π_u) with estimates $(\hat{Q}_u)$, write $b_u := \mathbb{E}[\hat{Q}_u \mid \mathcal{F}_u] - Q^{\pi_u}$ for the conditional bias and $\zeta_u := \hat{Q}_u - \mathbb{E}[\hat{Q}_u \mid \mathcal{F}_u]$ for the noise. Finally, $L_\delta := \ln(32|\mathcal{S}||\mathcal{A}|K/\delta)$.

Parameters. $\alpha := \sqrt{D_0+1}/(\bar{Q}\sqrt{\ln(eK)})$ and $\varsigma := (1-\gamma)^2\varepsilon/2$. Let $A_K := 9\sqrt{(D_0+1)L_\delta\ln(eK)}/(1-\gamma)^3$, and let K be the smallest integer with $K \ge 4A_K^2/\varepsilon^2$; it exists because the right-hand side grows polylogarithmically in K, and $K = \widetilde{O}(|\mathcal{A}|^{1-p}(1-\gamma)^{-6}\varepsilon^{-2})$ since $D_0 + 1 \le 2|\mathcal{A}|^{1-p}\log_2((1-\gamma)^{-1}) + 1$. The threshold $\underline{\pi}$ is defined in Lemma G.8 with $B_{\max} := 24(D_0+1)\sqrt{L_\delta}$.

Virtual process. Let $\sigma := \min\{t \ge 0 : \pi_t \notin \Pi_\theta\}$, with $\sigma = \infty$ if there is no such t; since π_t is $\mathcal{F}_t$-measurable, σ is a stopping time. The virtual process $(\check{\pi}_t)$ coincides with the algorithm up to time σ, i.e., $\check{\pi}_t = \pi_t$ for $t \le \sigma$. From time σ on, it performs the same mirror step with the exact $Q^{\check{\pi}_t}$ in place of the TOMC estimate. Its estimates are therefore $\check{Q}_t := \tilde{Q}^{\pi_t}$ for $t < \sigma$ and $\check{Q}_t := Q^{\check{\pi}_t}$ for $t \ge \sigma$, with bias $\check{b}_t$ and noise $\check{\zeta}_t$. The process is a device of the proof; the algorithm never uses exact values.

Lemma G.1 (Properties of the virtual process). *For every t and s: (a) $0 \le \check{Q}_t \le \bar{Q}$; (b) $\mathbb{E}[\check{\zeta}_t \mid \mathcal{F}_t] = 0$ and $\|\check{\zeta}_t\|_\infty \le \bar{Q}$; (c) $\check{b}_t(s,a) - \check{b}_t(s,a^*(s)) \ge -\varsigma$ for all a, hence $\langle \check{b}_t(s,\cdot), v - \pi^*(\cdot|s)\rangle \ge -\varsigma$ for every $v \in \Delta_{|\mathcal{A}|}$.*

Proof. (a) Observed costs lie in $[0,1]$, so TOMC estimates and exact Q-values lie in $[0,\bar{Q}]$. (b) $\check{\zeta}_t = \zeta_t \mathbf{1}\{t < \sigma\}$ with $\{t < \sigma\} \in \mathcal{F}_t$, and boundedness follows from (a). (c) For $t \ge \sigma$ the estimate is exact and $\check{b}_t = 0$. For $t < \sigma$, $\pi_t \in \Pi_\theta$, so $a^*(s)$ is non-rare at every state and every rare action differs from $a^*(s)$; Lemma 5.2 applies. □

Lemma G.2 (Cumulative form). *Let $\check{G}_t := \sum_{u<t} \eta_u \check{Q}_u$. For every t and s, $\check{\pi}_t(\cdot|s) = \arg\min_{v \in \Delta_{|\mathcal{A}|}} \{\langle \check{G}_t(s,\cdot), v\rangle + D^v_{\pi_0}(s)\}$; consequently $\langle \check{G}_t(s,\cdot), \check{\pi}_t(\cdot|s) - \pi^*(\cdot|s)\rangle \le D_0$.*

Proof. Write $\psi(v) := -\sum_a v_a^p/(p(1-p))$, so that $D^v_x = \psi(v) - \psi(x) - \langle \nabla\psi(x), v - x\rangle$ row by row. Since $\partial_a \psi(v) = -v_a^{p-1}/(1-p) \to -\infty$ as $v_a \to 0$ and $\check{Q}_u$ is bounded, every mirror step has an interior minimizer. It is characterized by $\eta_u \check{Q}_u(s,\cdot) + \nabla\psi(\check{\pi}_{u+1}(\cdot|s)) - \nabla\psi(\check{\pi}_u(\cdot|s)) + \lambda_u \mathbf{1} = 0$ for a scalar λ_u. Summing over $u < t$ gives $\check{G}_t(s,\cdot) + \nabla\psi(\check{\pi}_t(\cdot|s)) - \nabla\psi(\pi_0(\cdot|s)) + \Lambda_t \mathbf{1} = 0$, the optimality condition of the strictly convex problem in the statement. Comparing its objective at $\check{\pi}_t(\cdot|s)$ and at $\pi^*(\cdot|s)$ gives $\langle \check{G}_t(s,\cdot), \check{\pi}_t - \pi^*\rangle \le D^{\pi^*}_{\pi_0}(s) - D^{\check{\pi}_t}_{\pi_0}(s) \le D_0$. □

Lemma G.3 (Concentration). *Let $R_t := \bar{Q}\sqrt{8 S_t L_\delta}$ and $R'_t := \bar{Q}\sqrt{2 S_t L_\delta}$. With probability at least $1 - \delta/8$, for all $1 \le t \le K$, $s \in \mathcal{S}$ and $a \in \mathcal{A}$,*

$$\check{M}_t(s) := -\sum_{u<t} \eta_u \sum_q \kappa_s^{\pi^*}(q) \big\langle \check{\zeta}_u(q,\cdot), \check{\pi}_u(\cdot|q) - \pi^*(\cdot|q)\big\rangle \le R_t, \qquad \Big|\sum_{u<t} \eta_u \check{\zeta}_u(s,a)\Big| \le R'_t.$$

Denote this event by $\mathcal{E}_{\mathrm{opt}}$.

Proof. By Lemma G.1(b), and since $\check{\pi}_u$ is $\mathcal{F}_u$-measurable, both sums are martingales in t, with increments bounded by $2\eta_u\bar{Q}$ and $\eta_u\bar{Q}$ respectively. The Azuma–Hoeffding inequality bounds the failure probability by e^{-L_δ} for each (t,s) in the first case and by $2e^{-L_\delta}$ for each (t,s,a) in the second. A union bound over at most $3|\mathcal{S}||\mathcal{A}|K$ events, together with $e^{-L_\delta} = \delta/(32|\mathcal{S}||\mathcal{A}|K)$, gives the claim. □

The ergodic bound below applies the following one-step inequality at every state.

Lemma G.4 (Rowwise SPMD inequality). *For any state s, vector $g \in \mathbb{R}^{|\mathcal{A}|}$, current policy row x in the relative interior of $\Delta_{|\mathcal{A}|}$, comparator $y \in \Delta_{|\mathcal{A}|}$, and stepsize $\eta > 0$, the Tsallis-p mirror-descent update $x^+ := \arg\min_{v \in \Delta_{|\mathcal{A}|}} \{\langle g, v\rangle + \eta^{-1} D^v_x\}$ satisfies*

$$\eta\langle g, x - y\rangle \le D^y_x - D^y_{x^+} + \tfrac{\eta^2}{2}\|g\|^2_\infty.$$

Proof. The optimality condition and three-point Bregman identity give $\eta\langle g, x^+ - y\rangle \le D^y_x - D^y_{x^+} - D^{x^+}_x$. Adding $\eta\langle g, x - x^+\rangle \le \eta\|g\|_\infty\|x - x^+\|_1$ and using 1-strong convexity of the Tsallis divergence w.r.t. ℓ_1 ($D^{x^+}_x \ge \frac{1}{2}\|x - x^+\|_1^2$) plus Young's inequality yields the claim. □

Lemma G.5 (Ergodic bound and Bregman potential). *Let $\check{D}_t(s) := \sum_q \kappa_s^{\pi^*}(q) D^{\pi^*}_{\check{\pi}_t}(q)$. On $\mathcal{E}_{\mathrm{opt}}$, for all $1 \le t \le K$ and $s \in \mathcal{S}$,*

$$(1-\gamma)\sum_{u<t} \eta_u \big(V^{\check{\pi}_u}(s) - V^{\pi^*}(s)\big) + \check{D}_t(s) \le D_0 + \tfrac{1}{2}\bar{Q}^2 S_t + \varsigma W_t + R_t.$$

Proof. Apply Lemma G.4 at each state q, with $g = \check{Q}_u(q,\cdot)$ and comparator $\pi^*(\cdot|q)$:

$$\eta_u\langle \check{Q}_u(q,\cdot), \check{\pi}_u(\cdot|q) - \pi^*(\cdot|q)\rangle \le D^{\pi^*}_{\check{\pi}_u}(q) - D^{\pi^*}_{\check{\pi}_{u+1}}(q) + \tfrac{1}{2}\eta_u^2\bar{Q}^2.$$

Decompose $\check{Q}_u = Q^{\check{\pi}_u} + \check{b}_u + \check{\zeta}_u$. By Lemma G.1(c) with $v = \check{\pi}_u(\cdot|q)$, the bias term is at least $-\eta_u\varsigma$. Weight by $\kappa_s^{\pi^*}(q)$ and sum over q and $u < t$. The performance difference identity $\sum_q \kappa_s^{\pi^*}(q)\langle Q^\pi(q,\cdot), \pi(\cdot|q) - \pi^*(\cdot|q)\rangle = (1-\gamma)(V^\pi(s) - V^{\pi^*}(s))$, valid for every π, turns the first term into the value gap. The noise term is $\check{M}_t(s) \le R_t$, and $\check{D}_0(s) = D_0$. □

Lemma G.6 (Last-iterate bound). *On $\mathcal{E}_{\mathrm{opt}}$, for all $1 \le t \le K$,*

$$\|V^{\check{\pi}_t} - V^{\pi^*}\|_\infty \le \varepsilon_t := \frac{2D_0 + \frac{1}{2}\bar{Q}^2 S_t + R_t + 2R'_t}{(1-\gamma)^2 W_t} + \frac{\varsigma}{(1-\gamma)^2}.$$

Proof. Let $\bar{V}_t := W_t^{-1} \sum_{u<t} \eta_u V^{\check{\pi}_u}$ and $\bar{e}_t := \bar{V}_t - V^{\pi^*} \ge 0$. By linearity of the Bellman equation, $\widehat{Q}_t := W_t^{-1} \sum_{u<t} \eta_u Q^{\check{\pi}_u} = c + \gamma P \bar{V}_t = Q^{\pi^*} + \gamma P \bar{e}_t$. Decompose $\check{G}_t = W_t \widehat{Q}_t + \sum_{u<t} \eta_u \check{b}_u + \check{N}_t$, where $\check{N}_t := \sum_{u<t} \eta_u \check{\zeta}_u$. Lemma G.1(c) with $v = \check{\pi}_t(\cdot|s)$ bounds the bias part below by $-\varsigma W_t$, and $|\langle \check{N}_t(s,\cdot), \check{\pi}_t(\cdot|s) - \pi^*(\cdot|s)\rangle| \le 2\|\check{N}_t(s,\cdot)\|_\infty \le 2R'_t$. Lemma G.2 therefore yields

$$\langle \widehat{Q}_t(s,\cdot), \check{\pi}_t(\cdot|s) - \pi^*(\cdot|s)\rangle \le \varepsilon'_t := \frac{D_0 + 2R'_t}{W_t} + \varsigma.$$

Let $d_t(s) := \langle Q^{\pi^*}(s,\cdot), \check{\pi}_t(\cdot|s) - \pi^*(\cdot|s)\rangle \ge 0$. Since $\widehat{Q}_t = Q^{\pi^*} + \gamma P \bar{e}_t$ and $P_{\check{\pi}_t} \bar{e}_t \ge 0$, we have $d_t \le \varepsilon'_t \mathbf{1} + \gamma P_{\pi^*} \bar{e}_t$. The performance difference identity $V^{\check{\pi}_t} - V^{\pi^*} = (I - \gamma P_{\check{\pi}_t})^{-1} d_t$ then gives $\|V^{\check{\pi}_t} - V^{\pi^*}\|_\infty \le (\varepsilon'_t + \gamma \|\bar{e}_t\|_\infty)/(1-\gamma)$. Lemma G.5, after dropping the non-negative $\check{D}_t$, gives $\|\bar{e}_t\|_\infty \le (D_0 + \frac{1}{2}\bar{Q}^2 S_t + \varsigma W_t + R_t)/((1-\gamma)W_t)$. Combining the two, and using $(1-\gamma)(D_0 + 2R'_t) + \gamma(D_0 + \frac{1}{2}\bar{Q}^2 S_t + R_t) \le 2D_0 + \frac{1}{2}\bar{Q}^2 S_t + R_t + 2R'_t$ and $(1-\gamma)\varsigma + \gamma\varsigma = \varsigma$, yields the claim. □

Lemma G.7 (Parameter bounds). *For $1 \le t \le K$: $D_0 + \frac{1}{2}\bar{Q}^2 S_t + \varsigma W_t + R_t \le B_{\max}$ and $\varepsilon_t \le A_K/\sqrt{t} + \varepsilon/2$; in particular $\varepsilon_K \le \varepsilon$.*

Proof. From $S_t \le \alpha^2 \ln(et)$ and the choice of α: $\frac{1}{2}\bar{Q}^2 S_t \le \frac{1}{2}(D_0+1)$, $R_t \le \sqrt{8(D_0+1)L_\delta}$ and $R'_t \le \sqrt{2(D_0+1)L_\delta}$. Since $L_\delta \ge 1$, the numerator of ε_t is at most $9(D_0+1)\sqrt{L_\delta}$. Also $(1-\gamma)^2 W_t \ge (1-\gamma)^2 \alpha \sqrt{t} = (1-\gamma)^3 \sqrt{(D_0+1)t/\ln(eK)}$. These give the first term $A_K/\sqrt{t}$, while $\varsigma/(1-\gamma)^2 = \varepsilon/2$, and the choice of K gives $A_K/\sqrt{K} \le \varepsilon/2$. For the potential, minimality of K gives $\sqrt{K} < 2A_K/\varepsilon + 1$, so $\varsigma W_t \le 2\varsigma\alpha\sqrt{K} \le 19(D_0+1)\sqrt{L_\delta}$. Hence the potential bound is at most $\frac{3}{2}(D_0+1) + 19(D_0+1)\sqrt{L_\delta} + \sqrt{8(D_0+1)L_\delta} \le B_{\max}$. □

Lemma G.8 (Bregman-to-mass inversion). *Assume π^* is deterministic, and let $a^*(s)$ denote the action it selects at state s. If $(1-\gamma)D_\pi^{\pi^*}(s) \le B$, then*

$$\pi(a^*(s) \mid s) \ge \left(\frac{p}{1 + (1-p)p\, B/(1-\gamma)}\right)^{\frac{1}{1-p}}.$$

Consequently, with

$$\underline{\pi} := \min\left\{\frac{1}{|\mathcal{A}|},\ \left(\frac{p}{1 + (1-p)p\, B_{\max}/(1-\gamma)}\right)^{\frac{1}{1-p}}\right\},$$

$(1-\gamma)D_\pi^{\pi^*}(s) \le B_{\max}$ for all s implies $\pi \in \Pi_\theta$.

Proof. Let $x := \pi(a^*(s) \mid s)$. For Tsallis-p, deterministic π^* gives

$$\frac{-1 + p\, x^{p-1}}{(1-p)p} \le D_\pi^{\pi^*}(s).$$

Since $p - 1 < 0$, rearranging: $x^{p-1} \le (1 + (1-p)p\, D_\pi^{\pi^*}(s))/p$. Inverting (the map $x \mapsto x^{p-1}$ is decreasing) yields the claimed lower bound on x. □

Proposition G.9 (Part A). *On the event $\mathcal{E}_{\mathrm{opt}}$ of Lemma G.3, which has probability at least $1 - \delta/8$, $\check{\pi}_t = \pi_t \in \Pi_\theta$ for all $0 \le t \le K$, and $\|V^{\pi_t} - V^{\pi^*}\|_\infty \le \varepsilon_t \le A_K/\sqrt{t} + \varepsilon/2$ for all $1 \le t \le K$. In particular $V^{\pi_K}(s) - V^{\pi^*}(s) \le \varepsilon$ for all s.*

Proof. Work on $\mathcal{E}_{\mathrm{opt}}$. By Lemmas G.5 and G.7, $\check{D}_t(s) \le B_{\max}$, and $\kappa_s^{\pi^*}(s) \ge 1 - \gamma$ gives $(1-\gamma)D_{\check{\pi}_t}^{\pi^*}(s) \le B_{\max}$ for all s and $1 \le t \le K$. Lemma G.8 then gives $\check{\pi}_t \in \Pi_\theta$ for these t, and $\check{\pi}_0 = \pi_0 \in \Pi_\theta$ because $\underline{\pi} \le 1/|\mathcal{A}|$. If $\sigma \le K$, then $\pi_\sigma = \check{\pi}_\sigma \in \Pi_\theta$, contradicting the definition of σ; hence $\sigma > K$ and $\check{\pi}_t = \pi_t$ for all $t \le K$. The value bounds are Lemmas G.6 and G.7. □

Proof of Lemma 5.3. Work on $\mathcal{E}_{\mathrm{opt}}$, on which $\check{\pi}_t = \pi_t$ for all $t \le K$ (Proposition G.9). Write $D := D_0 + 1$, $\omega'(x) := -x^{p-1}/(1-p)$ for the derivative of the Tsallis potential (increasing on $(0,1]$), and $A^*(s,a) := Q^{\pi^*}(s,a) - V^{\pi^*}(s)$. Fix s, $1 \le t \le K$ and a with $\pi_t(a|s) \ge \underline{\pi}$, and let $\Psi_t := \omega'(\pi_t(a|s)) - \omega'(\pi_t(a^*(s)|s))$. Summing the optimality conditions as in the proof of Lemma G.2, and using that π_0 is uniform,

$$\Psi_t = -\sum_{u<t} \eta_u \big[\check{Q}_u(s,a) - \check{Q}_u(s,a^*(s))\big].$$

Decompose $\check{Q}_u = Q^{\check{\pi}_u} + \check{b}_u + \check{\zeta}_u$ and bound the three parts.

- Since $V^{\pi} \ge V^{\pi^*}$, every π satisfies $Q^{\pi}(s,a) - Q^{\pi}(s,a^*(s)) \ge A^*(s,a) - \gamma P(\cdot|s,a^*(s))^{\top}(V^{\pi} - V^{\pi^*})$. By Lemma G.5, $\sum_{u<t} \eta_u P(\cdot|s,a^*(s))^{\top}(V^{\check{\pi}_u} - V^{\pi^*}) \le (B_t^0 + \varsigma W_t)/(1-\gamma)$, where $B_t^0 := D_0 + \frac{1}{2}\bar{Q}^2 S_t + R_t$.
- Lemma G.1(c) gives $\check{b}_u(s,a) - \check{b}_u(s,a^*(s)) \ge -\varsigma$.
- Lemma G.3 bounds each of the two noise sums by R_t' in absolute value.

Hence

$$\Psi_t \le -W_t\big(A^*(s,a) - \varsigma\big) + \frac{\gamma}{1-\gamma}\big(B_t^0 + \varsigma W_t\big) + 2R_t'.$$

On the other hand, $\pi_t(a|s) \ge \underline{\pi}$ and $\pi_t(a^*(s)|s) \le 1$ give $\Psi_t \ge \omega'(\underline{\pi}) - \omega'(1) = -\Omega$, where $\Omega := (\underline{\pi}^{p-1} - 1)/(1-p)$. Combining the two bounds, and using $\varsigma + \gamma\varsigma/(1-\gamma) = \varsigma/(1-\gamma)$,

$$A^*(s,a) \le \frac{\varsigma}{1-\gamma} + \frac{\Omega + 2R_t' + \gamma B_t^0/(1-\gamma)}{W_t}.$$

It remains to bound Ω. If $\underline{\pi}$ equals the second term in its definition (Lemma G.8), then $\Omega = 1/p + B_{\max}/(1-\gamma)$ exactly. Otherwise $\underline{\pi} = 1/|\mathcal{A}|$, and $\Omega = pD_0 \le B_{\max}/(1-\gamma)$. In both cases $\Omega \le 1/p + B_{\max}/(1-\gamma)$. Moreover $1/p = \log_2((1-\gamma)^{-1}) \le \frac{1}{2(1-\gamma)}$ for $\gamma \ge 3/4$. With the bounds of Lemma G.7 ($B_t^0 \le \frac{3}{2}D + \sqrt{8DL_\delta}$ and $R_t' \le \sqrt{2DL_\delta}$), the numerator is at most $32D\sqrt{L_\delta}/(1-\gamma)$. Since $W_t \ge \alpha\sqrt{t} = (1-\gamma)\sqrt{Dt/\ln(eK)}$, the second term is at most $32\sqrt{DL_\delta \ln(eK)}/((1-\gamma)^2\sqrt{t}) \le 4(1-\gamma)A_K/\sqrt{t}$, and $\varsigma/(1-\gamma) = (1-\gamma)\varepsilon/2$. Also $A^*(s,a) \le (1-\gamma)^{-1}$ always, so $A^*(s,a) \le u_t$. Finally, if $s \in \mathcal{S}_t^{\mathrm{multi}}$, some $a \ne a^*(s)$ is non-rare, and $\Delta(s) \le A^*(s,a) \le u_t$. □

H Sample complexity proof (Part B of Theorem 3.7)

The goal is to bound $\sum_{t=0}^{K-1} \tilde{m}_t^{\mathrm{loc}}$. Let $A_t := \{\pi_t \in \Pi_\theta\}$. Part B works on the event $\mathcal{E}_{\mathrm{opt}}$ of Lemma G.3, on which every A_t with $t \le K$ holds (Proposition G.9) and the coordinate locking of Lemma 5.3 applies. The argument has four steps: (1) turn the per-iteration hitting cost into a high-probability event; (2) bound it through H_θ; (3) shrink the target set by locking; (4) decompose into phases.

Step 1: Per-iteration cost (high-probability event). For a stationary policy π and a pair (s,a), define the high-probability hitting time

$$\mathrm{Hit}_{\pi,\delta',\varsigma}(s,a) := \min\big\{m \ge 1 : \Pr\big(h_m^{\pi}(s,a) \le \varsigma(1-\gamma) \,\big|\, s_0\big) \ge 1-\delta',\ \forall s_0 \in \mathcal{S}\big\},$$

the smallest rollout length that drives the discounted hitting residual below $\varsigma(1-\gamma)$ with probability at least $1-\delta'$ from every starting state, and the residual-decay length

$$L_\gamma(\varsigma) := \left\lceil \frac{\log(1/(\varsigma(1-\gamma)))}{|\log\gamma|} \right\rceil \le \frac{1}{1-\gamma}\ln\frac{1}{\varsigma(1-\gamma)} + 1, \tag{6}$$

where the inequality uses $|\ln\gamma| \ge 1-\gamma$. Let $\overline{\mathrm{Hit}}(s) := L_\gamma(\varsigma) + \lceil 2\theta^{-1}H_\theta(s)\ln(e/\delta')\rceil$ be the envelope of Lemma C.1, and set $\delta' := \delta/(2|\mathcal{Z}|K)$. For each t and $(s,a) \in \mathcal{Z}$, let $B_{t,s,a} := \{s \in \mathcal{S}_t^{\mathrm{multi}},\ \pi_t(a|s) \ge \theta\}$ and $\mathcal{E}_{t,s,a} := \{h^{\pi_t}_{\overline{\mathrm{Hit}}(s)}(s,a) \le \varsigma(1-\gamma)\}$. The events A_t and $B_{t,s,a}$ are $\mathcal{F}_t$-measurable, and on $A_t \cap B_{t,s,a}$ Lemma C.1 applies to π_t. Since h_m is non-increasing in m once the pair has been visited,

$$\mathbb{E}\big[\mathbf{1}_{A_t}\mathbf{1}_{B_{t,s,a}}\mathbf{1}_{\mathcal{E}_{t,s,a}^c} \,\big|\, \mathcal{F}_t\big] \le \delta'\,\mathbf{1}_{A_t}\mathbf{1}_{B_{t,s,a}}.$$

Taking expectations and a union bound over $t \in \{0, \dots, K-1\}$ and $(s,a) \in \mathcal{Z}$ shows that, with probability at least $1 - \delta/2$, $\mathcal{E}_{t,s,a}$ holds for every (t,s,a) with $A_t \cap B_{t,s,a}$. Intersecting with $\mathcal{E}_{\mathrm{opt}}$ (probability at least $1 - \delta/8$), on which every A_t holds, we obtain an event of probability at least $1 - \delta/2 - \delta/8 \geq 1 - \delta$ on which, for all t and with $\max \varnothing := 0$,

$$\tilde{m}_t^{\mathrm{loc}} \leq 1 + \max_{s \in \mathcal{S}_t^{\mathrm{multi}}} \overline{\mathrm{Hit}}(s).$$

Certification enters only through the $\mathcal{F}_t$-measurable indicator $\mathbf{1}_{A_t}$, so no stopped process is needed in Part B.

Step 2: Certified hitting envelope. Since $H_\theta(s) \geq 1$, $\lceil x \rceil \leq 2x$ for $x = 2\theta^{-1} H_\theta(s) \ln(e/\delta') \geq 1$, and the envelope is algorithm-independent:

$$\overline{\mathrm{Hit}}(s) \leq L_\gamma(\varsigma) + \frac{4}{\theta} \ln \frac{e}{\delta'} \cdot H_\theta(s).$$

Only $\ln(e/\delta')$ appears; no parameter of the algorithm depends on H_θ.

Step 3: Active-set shrinkage via coordinate locking. The event of Step 1 is contained in $\mathcal{E}_{\mathrm{opt}}$, on which Lemma 5.3 gives $\mathcal{S}_t^{\mathrm{multi}} \subseteq \{s : \Delta(s) \leq u_t\}$ for $1 \leq t \leq K$. With $u_0 := (1-\gamma)^{-1}$ the containment is trivial at $t = 0$. Hence

$$\tilde{m}_t^{\mathrm{loc}} \leq 1 + L_\gamma(\varsigma) + \frac{4}{\theta} \ln \frac{e}{\delta'} \cdot \max_{\Delta(s) \leq u_t} H_\theta(s).$$

Step 4: Phase decomposition. Write $A^{\mathrm{lk}} := 4(1-\gamma)A_K$ and $\varepsilon^{\mathrm{lk}} := 4(1-\gamma)\varepsilon$. Then $u_t = \min\{(1-\gamma)^{-1}, A^{\mathrm{lk}}/\sqrt{t} + \varepsilon^{\mathrm{lk}}/2\}$ and $4(A^{\mathrm{lk}})^2/(\varepsilon^{\mathrm{lk}})^2 = 4A_K^2/\varepsilon^2$. Let $J^{\mathrm{lk}} := \lceil \log_2(1/\varepsilon^{\mathrm{lk}}) \rceil$ and $I_j := \{t < K : u_t \in (2^{-(j+1)}, 2^{-j}]\}$. Since $u_t \leq (1-\gamma)^{-1} \leq 2^{-j_{\min}}$ and $u_t \geq \varepsilon^{\mathrm{lk}}/2 \geq 2^{-(J^{\mathrm{lk}}+1)}$, every $t < K$ lies in some I_j with $j_{\min} \leq j \leq J^{\mathrm{lk}} + 1$.

- If $2^{-(j+1)} \geq \varepsilon^{\mathrm{lk}}$, then every $t \in I_j$ with $t \geq 1$ satisfies $A^{\mathrm{lk}}/\sqrt{t} \geq u_t - \varepsilon^{\mathrm{lk}}/2 > 2^{-(j+1)} - \varepsilon^{\mathrm{lk}}/2 \geq 2^{-(j+2)}$, so $t < 16(A^{\mathrm{lk}})^2\, 4^j$. The index $t = 0$ adds at most one element, so $|I_j| \leq 16(A^{\mathrm{lk}})^2\, 4^j + 1$.
- Otherwise $j \geq J^{\mathrm{lk}} - 1$, and the minimality of K gives $|I_j| \leq K < 4(A^{\mathrm{lk}})^2/(\varepsilon^{\mathrm{lk}})^2 + 1 \leq 16(A^{\mathrm{lk}})^2\, 4^j + 1$.

Hence $|I_j| \leq 16(A^{\mathrm{lk}})^2\, 4^j + 1 = \widetilde{O}(|\mathcal{A}|^{1-p}(1-\gamma)^{-4}\, 4^j)$ for every j. In phase j the cost beyond the baseline $1 + L_\gamma(\varsigma)$ is at most $\frac{4}{\theta} \ln \frac{e}{\delta'} M_j$, where $M_j := \max_{\Delta(s) \leq 2^{-j}} H_\theta(s)$ is non-increasing in j.

Summation. Let $N_{\mathrm{ph}} := J^{\mathrm{lk}} - j_{\min} + 2$ be the number of phases, and let $\mathcal{I}_{\mathrm{ph}}(H_\theta; \varepsilon^{\mathrm{lk}})$ be (4) with ε replaced by $\varepsilon^{\mathrm{lk}}$, a sum over $j_{\min} \leq j \leq J^{\mathrm{lk}}$. Since M_j is non-increasing, the extra phase satisfies $4^{J^{\mathrm{lk}}+1} M_{J^{\mathrm{lk}}+1} \leq 4 \cdot 4^{J^{\mathrm{lk}}} M_{J^{\mathrm{lk}}}$, so $\sum_{j=j_{\min}}^{J^{\mathrm{lk}}+1} 4^j M_j \leq 5\, \mathcal{I}_{\mathrm{ph}}(H_\theta; \varepsilon^{\mathrm{lk}})$. The terms $+1$ contribute at most $N_{\mathrm{ph}} M_{j_{\min}}$, and $M_{j_{\min}} \leq 4^{-j_{\min}} \mathcal{I}_{\mathrm{ph}}(H_\theta; \varepsilon^{\mathrm{lk}}) \leq 4(1-\gamma)^{-2} \mathcal{I}_{\mathrm{ph}}(H_\theta; \varepsilon^{\mathrm{lk}})$. Hence

$$\sum_{j=j_{\min}}^{J^{\mathrm{lk}}+1} |I_j|\, M_j \leq \left[80(A^{\mathrm{lk}})^2 + 4N_{\mathrm{ph}}(1-\gamma)^{-2}\right] \mathcal{I}_{\mathrm{ph}}(H_\theta; \varepsilon^{\mathrm{lk}}),$$

and summing Step 3 over $t < K$,

$$\begin{aligned} \sum_{t=0}^{K-1} \tilde{m}_t^{\mathrm{loc}} &\leq K\big(1 + L_\gamma(\varsigma)\big) + \frac{4}{\theta} \ln \frac{e}{\delta'} \sum_{j=j_{\min}}^{J^{\mathrm{lk}}+1} |I_j|\, M_j \\ &\leq K\big(1 + L_\gamma(\varsigma)\big) + \widetilde{O}\left(\frac{(1-\gamma)^2 A_K^2}{\theta}\, \mathcal{I}_{\mathrm{ph}}(H_\theta; \varepsilon^{\mathrm{lk}}) \right), \end{aligned} \tag{7}$$

because $80(A^{\mathrm{lk}})^2 = 1280(1-\gamma)^2 A_K^2$ and $4(1-\gamma)^{-2} \leq (1-\gamma)^2 A_K^2$, while N_{ph} and $\ln(e/\delta')$ are logarithmic. Since $(1-\gamma)^2 A_K^2 = \widetilde{O}(|\mathcal{A}|^{1-p}(1-\gamma)^{-4})$ and $\mathcal{I}_{\mathrm{ph}}(H_\theta; \varepsilon^{\mathrm{lk}}) \leq \mathcal{I}_{\mathrm{ph}}(H_\theta; (1-\gamma)\varepsilon)$ (the second sum has more phases), this is (5); Proposition 3.5 gives the $\mathcal{I}_\infty$ form. The baseline cannot be absorbed: when $\Delta(s) = \Theta(1)$ for all s, $\mathcal{I}_{\mathrm{ph}}(H_\theta; (1-\gamma)\varepsilon) = O(\max_s H_\theta(s))$, while $K = \widetilde{\Theta}(\varepsilon^{-2})$.

I LOWER BOUND PROOFS

This section proves Propositions 4.4 and 4.5. Divergences are in nats. An algorithm is a sequence of action-selection kernels, each mapping the history (states, actions, costs) to a distribution over actions, together with an output rule; $\mathbb{P}_i^T$ is the law of the history of T transitions under $\mathcal{M}_i$. We write $N_s(T) := \sum_{t=0}^{T-1} \mathbf{1}\{S_t = s\}$ and $N_{s,a}(T)$ for the analogous pair count. In both families, $\mathcal{M}_1$ satisfies the conclusions of Lemma 4.2 with the roles of L and R exchanged at one state: the relabeling maps the optimal policy of $\mathcal{M}_0$ to an optimal policy of $\mathcal{M}_1$ with the same state chain, the same values and the same gaps.

Lemma I.1 (Reduction to testing)**.** *Let $\hat{\pi}$ be a stationary policy and s a state with $\Delta(s) > 0$. If $V^{\hat{\pi}}(s) - V^{\pi^*}(s) \le \varepsilon$, then $1 - \hat{\pi}(a^*(s)|s) \le \varepsilon/\Delta(s)$.*

Proof. This is Lemma D.1 with $\pi = \hat{\pi}$ and $\varepsilon' = \varepsilon$. □

Lemma I.2 (Occupancy)**.** *On the ring, for every algorithm and every $T \ge 1$, the number of entrances $J_b(T) := \sum_{t=0}^{T-1} \mathbf{1}\{S_t \ne b,\ S_{t+1} = b\}$ satisfies $\mathbb{E}[J_b(T)] \le rT$. Starting from g_2, $N_{g_1}(T) \le T/(2(n-2)) + J_b(T)$ pathwise; hence $\mathbb{E}[N_{g_1}(T)] \le T\big(\frac{1}{2(n-2)} + \frac{1}{n^2}\big) \le \frac{5T}{4n}$ for $n \ge 4$.*

Proof. From every state other than b the chain enters b with probability exactly r, whatever action is taken, so $\mathbb{E}[J_b(T)] = r\,\mathbb{E}[\sum_{t<T} \mathbf{1}\{S_t \ne b\}] \le rT$. A visit to g_1 is entered either from b or from $\{x_m, y_m\}$. Each sojourn at b begins with an entrance and ends with at most one move to g_1, and the chain starts at $g_2 \ne b$, so there are at most $J_b(T)$ visits of the first kind. A visit of the second kind ends a jump-free passage along the ring that starts at the previous visit to g_1 or at the initial state g_2; it lasts $2(n-1)$ steps in the first case and $2(n-2)$ steps in the second, and the passages of distinct visits are disjoint intervals of time. Hence there are at most $T/(2(n-2))$ visits of the second kind. Finally, $\frac{1}{2(n-2)} + \frac{1}{n^2} \le \frac{5}{4n}$ for $n \ge 4$. □

Lemma I.3 (Divergence decomposition)**.** *Let $\mathcal{M}_0, \mathcal{M}_1$ share $\mathcal{S}$, $\mathcal{A}$, the initial state and the deterministic cost function, and let P_0, P_1 be their kernels. For every algorithm,*

$$\mathrm{KL}(\mathbb{P}_0^T \| \mathbb{P}_1^T) = \sum_{(s,a) \in \mathcal{Z}} \mathbb{E}_0[N_{s,a}(T)]\, \mathrm{KL}\big(P_0(\cdot|s,a) \,\|\, P_1(\cdot|s,a)\big).$$

Proof. Apply the chain rule to the history $(S_0, A_0, S_1, A_1, \ldots, S_T)$. The conditional law of each action given the past is the same under both instances (same algorithm, same history), and the costs are the same deterministic function of (S_t, A_t), so these terms vanish. The term of S_{t+1} given the past is $\mathrm{KL}(P_0(\cdot|S_t, A_t) \| P_1(\cdot|S_t, A_t))$; taking $\mathbb{E}_0$ and summing over $t < T$ gives the claim. □

Proof of Proposition 4.4. In $\mathcal{M}_1$ the action R at b moves to g_1, while L stays or moves with probability $\frac{1}{2}$ each, so $a_1^*(b) = R$ and $a_0^*(b) = L$; both instances have $\Delta(b) \ge 45/256 > 1/6$ (Lemma 4.2(i)). By Lemma I.1, an ε-optimal output with $\varepsilon < 1/12$ has $\hat{\pi}(a_i^*(b)|b) > \frac{1}{2}$ under $\mathcal{M}_i$. The test $\psi := \mathbf{1}\{\hat{\pi}(R|b) > \frac{1}{2}\}$ therefore satisfies $\mathbb{P}_0^T(\psi = 1) \le \frac{1}{3}$ and $\mathbb{P}_1^T(\psi = 1) \ge \frac{2}{3}$, so $\mathrm{TV}(\mathbb{P}_0^T, \mathbb{P}_1^T) \ge \frac{1}{3}$. The two instances differ only in the kernel at b and start at $g_2 \ne b$, so their histories can be coupled to coincide until the first transition out of b. Hence $\mathrm{TV}(\mathbb{P}_0^T, \mathbb{P}_1^T) \le \Pr(\tau_b < T) \le \mathbb{E}[J_b(T)] \le rT$ by Lemma I.2, and $T \ge 1/(3r) = n^2/3$. □

Proof of Proposition 4.5. With $\eta := 3\varepsilon/(2\gamma a_r)$, Lemma 4.2(i) gives $\Delta(g_1) = 3\varepsilon$ in both instances, and $\gamma \ge 3/4$, $n \ge 4$ give $\eta \le \frac{32}{15}\varepsilon < \frac{1}{4}$ for $\varepsilon < 1/12$. In $\mathcal{M}_1$ the roles of L and R at g_1 are exchanged, so $a_1^*(g_1) = R$ and $a_0^*(g_1) = L$. By Lemma I.1, success forces $\hat{\pi}(a_i^*(g_1)|g_1) \ge \frac{2}{3}$, so the test $\mathbf{1}\{\hat{\pi}(R|g_1) \ge \frac{2}{3}\}$ gives $\mathrm{TV}(\mathbb{P}_0^T, \mathbb{P}_1^T) \ge \frac{1}{3}$, and Pinsker's inequality gives $\mathrm{KL}(\mathbb{P}_0^T \| \mathbb{P}_1^T) \ge 2\,\mathrm{TV}^2 \ge \frac{2}{9}$. The kernels differ only at g_1, where under either action the next state is b with probability r in both instances, and otherwise y_1 or x_1 with the probabilities $\frac{1}{2} \pm \eta$ exchanged. By Lemma I.3,

$$\mathrm{KL}(\mathbb{P}_0^T \| \mathbb{P}_1^T) = a_r \cdot 2\eta \ln \frac{1+2\eta}{1-2\eta} \cdot \mathbb{E}_0[N_{g_1}(T)] \le 16\eta^2 \cdot \frac{5T}{4n},$$

using $\ln \frac{1+x}{1-x} \le \frac{2x}{1-x^2} \le \frac{8x}{3}$ for $x = 2\eta \le \frac{1}{2}$ and Lemma I.2. Combining, $20\eta^2 T/n \ge \frac{2}{9}$, i.e., $T \ge n/(90\eta^2)$. □

J AN ACTION-DEPENDENT STAR

The star is the intermediate case of Section 3.5: only the bottleneck has an action-dependent kernel, and H_θ stays within a fixed power of θ (Lemma J.2).

Example J.1 (Action-dependent star)**.** Let $n \geq 4$ and $\varepsilon \in (0, 1/32]$. The states are a hub h, a bottleneck b and generic states $g_1, \dots, g_{n-2}$, and the actions are $\{0, 1\}$. Under action 0, $P_0(h,h) = \frac{1}{2}$, $P_0(h,b) = \frac{1}{2n^2}$, $P_0(h,g_i) = \frac{1-n^{-2}}{2(n-2)}$, and every non-hub state moves to h. Action 1 agrees with action 0 except at b, where $P_1(b,b) = 1$. With $v(s) := \frac{1}{2(1-\gamma)} + \frac{1}{16}\mathbf{1}\{s = b\}$, $d(h) = d(b) := \frac{1}{32}$ and $d(g_i) := \varepsilon$, the costs are $c(s,0) := v(s) - \gamma(P_0 v)(s)$ and $c(s,1) := v(s) - \gamma(P_1 v)(s) + d(s)$.

At b the worse action has the lower immediate cost, $c(b,1) - c(b,0) = \frac{1}{32} - \frac{\gamma}{16} < 0$, while $Q^{\pi^*}(b,1) - V^{\pi^*}(b) = \frac{1}{32}$: the optimal action cannot be read off from immediate costs.

Lemma J.2 (Star parameters)**.** *For $\gamma \in [3/4, 1)$, the costs lie in $[0,1]$, π^* plays action 0 everywhere, $V^{\pi^*} = v$, and $\Delta = d$. Only b has an action-dependent kernel, and for every $\theta \in (0,1]$: $H_\theta(h) \leq 1/\theta$ and $H_\theta(s) \leq 2/(\theta P_0(h,s))$ for $s \neq h$.*

Proof. Costs. $c(h,0) = \frac{1}{2} - \frac{\gamma}{32n^2}$, $c(g_i,0) = \frac{1}{2}$, $c(b,0) = \frac{1}{2} + \frac{1}{16}$, $c(b,1) = \frac{1}{2} + \frac{1-\gamma}{16} + \frac{1}{32}$, and $c(s,1) = c(s,0) + d(s)$ for $s \in \{h, g_i\}$; all lie in $[0,1]$. *Optimality.* By construction $c(s,0) + \gamma(P_0 v)(s) = v(s)$ and $c(s,1) + \gamma(P_1 v)(s) = v(s) + d(s)$, so v is the fixed point of the Bellman optimality operator, action 0 is optimal everywhere, and $\Delta(s) = d(s)$. *Hitting.* Let $\pi \in \Pi_\theta$. From any $x \neq h$ the chain moves to h in one step with probability at least θ (probability 1 unless $x = b$), so $q_h^\theta(1) \geq \theta$. For $s \neq h$ and any start, the chain is at h after at most one step with probability at least θ, and the kernel at h does not depend on the action, so it then moves to s with probability $P_0(h,s)$; hence $q_s^\theta(2) \geq \theta P_0(h,s)$. □

Corollary J.3 (Star bound)**.** *On Example J.1 with γ fixed, the exploration term of Theorem 3.7 is $\widetilde{O}(\theta^{-2}(n^2 + n\varepsilon^{-2})) = \widetilde{O}(n^2 + n\varepsilon^{-2})$, although $\mathcal{S}_{\mathrm{dep}} = \{b\} \neq \varnothing$.*

Proof. By Lemma J.2, $H_\theta(b) \leq 4n^2/\theta$, $H_\theta(g_i) \leq 4(n-2)/(\theta(1-n^{-2})) = O(n/\theta)$ and $H_\theta(h) \leq 1/\theta$. The states h and b have gap $\frac{1}{32}$ and are unresolved only in phases with $2^{-j} \geq \frac{1}{32}$, contributing $O(n^2/\theta)$ to $\mathcal{I}_{\mathrm{ph}}(H_\theta; \varepsilon')$ for every accuracy level $\varepsilon' \in (0,1)$; the generic states contribute $O(n\varepsilon'^{-2}/\theta)$. Theorem 3.7 with $\varepsilon' = (1-\gamma)\varepsilon$, together with $\theta^{-1} = \widetilde{O}(1)$ for fixed γ and $|\mathcal{A}| = 2$, concludes. □

For comparison, every row of the optimal kernel puts mass at least $\frac{1}{2}$ on h. Hence the optimal chain satisfies Assumption 3.10 with $C^* = 1$ and $\rho^* = \frac{1}{2}$, and it has $\nu^*(h) = \frac{2}{3}$, $\nu^*(b) = \frac{1}{3n^2}$ and $\nu^*(g_i) = \frac{1-n^{-2}}{3(n-2)}$. So $\bar{b} = \lceil \log_2(12n^2) \rceil = O(\log n)$, and the mixing regime of Corollary 3.11 pays $\theta^{-O(\log n)}$. At the same hitting cost, the global charge is at least $\varepsilon^{-2} H_\theta(b) \geq 2n^2\varepsilon^{-2}$, because $H_\theta(b) \geq D_\theta(b) \geq \mathbb{E}_h^{\pi^*}[\tau_b] \geq 2n^2$. The separation is therefore $\widetilde{\Theta}(\min\{n, \varepsilon^{-2}\})$ for fixed γ, as on the ring.

K PROOFS FOR THE RING (LEMMA 4.2, PROPOSITION 4.3) AND ITS SCOPE

Proof of Lemma 4.2. (i) Optimality and gaps. The states x_i and y_i have the same successors under both actions and costs 1 and 0, so $V^\pi(x_i) - V^\pi(y_i) = 1$ for every stationary π; their two actions coincide, so $\Delta(x_i) = \Delta(y_i) = 0$. At g_i the two actions share the cost and the jump probability, so for every π

$$Q^\pi(g_i, R) - Q^\pi(g_i, L) = \gamma a_r (\zeta_L - \zeta_R)\big(V^\pi(x_i) - V^\pi(y_i)\big) = 2\gamma a_r \eta.$$

At b, $Q^\pi(b,R) - Q^\pi(b,L) = \frac{\gamma}{2}(V^\pi(b) - V^\pi(g_1))$. Let π^L play L everywhere. Under π^L, a state of cost 1 (b or some x_i) is followed by a state of cost 1 with probability at most r, so $\mathbb{E}[c_t + c_{t+1}] \leq 1 + r$ for all t. Starting from g_1 we have $c_0 = 0$, and pairing the terms $(1,2), (3,4), \dots$ gives $V^{\pi^L}(g_1) \leq \gamma(1+r)/(1-\gamma^2)$. Since $V^{\pi^L}(b) = 1 + \gamma V^{\pi^L}(g_1)$,

$$V^{\pi^L}(b) - V^{\pi^L}(g_1) = 1 - (1-\gamma)V^{\pi^L}(g_1) \geq \frac{1-\gamma r}{1+\gamma} \geq \frac{a_r}{2}.$$

Both Q-differences are positive under π^L, so π^L admits no improving action and is optimal. At $\pi^* = \pi^L$ the displayed identities give $\Delta(g_i) = 2\gamma a_r \eta$ and $\gamma a_r/4 \le \Delta(b) \le \gamma/2$, the upper bound because $V^{\pi^L}(b) - V^{\pi^L}(g_1) \le 1$. For $\gamma \ge 3/4$ and $n \ge 4$, $\gamma a_r/4 \ge \frac{3}{4} \cdot \frac{15}{16} \cdot \frac{1}{4} = \frac{45}{256}$.

(ii) Stationary distribution and mixing. Under π^*, b is entered from every other state with probability r and left at once, so $\nu^*(b) = r(1-\nu^*(b))$, i.e., $\nu^*(b) = r/(1+r)$. Flow balance along the ring gives $\nu^*(x_i) + \nu^*(y_i) = a_r \nu^*(g_i)$, $\nu^*(g_{i+1}) = a_r^2 \nu^*(g_i)$ for $i < m$, and $\nu^*(g_1) = \nu^*(b) + a_r^{2m}\nu^*(g_1)$; hence

$$\nu^*(g_i) = \frac{r}{1+r} \cdot \frac{a_r^{2(i-1)}}{1-a_r^{2m}}, \qquad \nu^*(x_i) = a_r\big(\tfrac{1}{2} - \eta\big)\nu^*(g_i), \qquad \nu^*(y_i) = a_r\big(\tfrac{1}{2} + \eta\big)\nu^*(g_i).$$

Since $mr \le 1/n$, we have $1 - a_r^{2m} \in [2mr(1-mr), 2mr]$ and $a_r^{2m} \ge \frac{1}{2}$; with $\eta \le \frac{1}{4}$, all of these are $\Theta(1/n)$. For mixing, $P^{(2)}_{\pi^*}(b,b) = r$ and $P^{(2)}_{\pi^*}(z,b) = a_r r$ for $z \ne b$ (stay off b, then jump), so $P^{(2)}_{\pi^*}(z,\cdot) \ge a_r r\, \delta_b$ for every z. This two-step Doeblin minorization gives $\|P^{(t)}_{\pi^*}(z,\cdot) - \nu^*\|_{\mathrm{tv}} \le (1 - a_r r)^{\lfloor t/2 \rfloor} \le 2(\sqrt{1-a_r r})^t$, i.e., $C^* = 2$ and $\rho^* = \sqrt{1-a_r r} \in [1/2, 1)$. The mixing is genuinely slow. Group the states other than b into the $2m$ ring positions $\{g_i\}$ and $\{x_i, y_i\}$; each has stationary mass at most $\nu^*(g_1) \le 1/(2m(1-mr)) \le \frac{2}{9}$ for $n \ge 4$. From $z \ne b$, the chain makes no jump to b in t steps with probability a_r^t, and on that event it lies at time t in a position C_t determined by z and t. For $t = \lfloor n^2/2 \rfloor$, $a_r^t \ge 1 - tr \ge \frac{1}{2}$, so $\|P^{(t)}_{\pi^*}(z,\cdot) - \nu^*\|_{\mathrm{tv}} \ge P^{(t)}_{\pi^*}(z, C_t) - \nu^*(C_t) \ge \frac{1}{2} - \frac{2}{9} > \frac{1}{4}$, and the mixing time at tolerance $\frac{1}{4}$ exceeds $\lfloor n^2/2 \rfloor$. Irreducibility is immediate, and $\underline{\nu}^* = \Theta(n^{-2})$ together with $\log(1/\rho^*) = \Theta(n^{-2})$ gives $\bar{b} = \Theta(n^2 \log n)$.

(iii) Hitting costs. From every $x \ne b$ the chain enters b at the next step with probability exactly r under any action, so $q_b^\theta(\ell) = 1 - (1-r)^\ell \le \ell r$ for every θ, with equality at $\ell = 1$; hence $H_\theta(b) = \inf_\ell \ell/(1-(1-r)^\ell) = 1/r = n^2$. Now let $s \ne b$. The following bounds hold for every policy and every history:

- from any ring state, s is reached within $2m$ further transitions if no jump occurs and the chain takes the required branch at the last branch state before s;
- from b, the chain leaves within two steps with probability at least $\frac{3}{4}$, and it is then at g_1, at most $2m-1$ transitions before s;
- no jump occurs during $2n$ transitions with probability at least $(1-r)^{2n} \ge 1 - 2/n \ge \frac{1}{2}$;
- each branch is taken with conditional probability at least $\frac{1}{2} - \eta \ge \frac{1}{4}$.

Multiplying these conditional bounds gives $q_s^\theta(2n) \ge \frac{3}{4} \cdot \frac{1}{2} \cdot \frac{1}{4} = \frac{3}{32}$, so $H_\theta(s) \le 2n \cdot \frac{32}{3} = \frac{64n}{3}$. For the lower bound, start at the ring position just after that of s: at x_i when $s = g_i$, and at g_{i+1} when $s \in \{x_i, y_i\}$. Without a jump, reaching s takes at least m steps, and no jump occurs in the first m steps with probability $(1-r)^m \ge 1 - 1/n \ge \frac{3}{4}$. Hence $\mathbb{E}_x^{\pi^*}[\tau_s] \ge \frac{3}{4}m \ge \frac{n}{2}$, and Proposition 3.2 gives $H_\theta(s) \ge D_\theta(s) \ge \frac{n}{2}$. □

Proof of Proposition 4.3. By Lemma 4.2, uniformly in θ: $H_\theta(b) = n^2$, $H_\theta(s) \le 64n/3$ for $s \ne b$, and $\Delta(b) \ge 45/256$. The bottleneck is unresolved only in phases with $2^{-j} \ge 45/256$, i.e., $j \le 2$. So for every accuracy level $\varepsilon' \in (0,1)$, with $J' := \lceil \log_2(1/\varepsilon') \rceil$,

$$\mathcal{I}_{\mathrm{ph}}(H_\theta; \varepsilon') \le \sum_{j=j_{\min}}^{2} 4^j n^2 + \sum_{j=j_{\min}}^{J'} 4^j \cdot \frac{64n}{3} \le \frac{64}{3} n^2 + \frac{256}{9} \cdot 4^{J'} n \le 22n^2 + 114\, n\varepsilon'^{-2}.$$

For $|\mathcal{A}| = 2$ and fixed γ, $\theta^{-1} = \widetilde{O}(1)$ (Lemma G.8), and Theorem 3.7, which uses $\varepsilon' = (1-\gamma)\varepsilon$, gives exploration term $\widetilde{O}(n^2 + n\varepsilon^{-2})$. □

Remark K.1 (What the family does and does not show)**.** The layer x_i, y_i is what makes a deterministic-cost lower bound match. It places the perturbation on a branch of probability $\approx \frac{1}{2}$ whose outcomes differ in value by exactly 1; perturbing the jump probability r instead also gives gap $\Theta(\eta)$, but per-visit information $\Theta(\eta^2/r)$ and a lower bound smaller by a factor n^2. The two actions at g_i differ in kernel by $4\eta a_r = \Theta(\varepsilon)$. This is a feature of the symmetric relabeling construction, not of the lower-bound method. In a two-point argument of order ε^{-2}, Pinsker's inequality constrains the perturbation between the two instances at the distinguished pair to $O(\varepsilon)$, but it does not constrain

the difference between two actions within an instance. For example, over three successors with values $0, \frac{1}{2}, 1$, the kernels $(\frac{5}{12} \pm h, \frac{1}{6}, \frac{5}{12} \mp h)$ and $(\frac{1}{6}, \frac{2}{3}, \frac{1}{6})$ are at ℓ_1-distance 1, while their expected values differ by h and the two signs are at divergence $\Theta(h^2)$. The tied states x_i, y_i stay unresolved throughout but are charged only their own $H_\theta = \Theta(n)$: a zero gap costs the reachability of the tied state, not the global charge. The family isolates short-range branching; it is not a model of long-horizon credit assignment.

L Proof of Proposition 3.5

Proof. Upper bound. For each j with a nonempty inner set, pick s_j attaining $\max_{\Delta(s)\le 2^{-j}} w(s)$. Then $4^j \le \Delta(s_j)^{-2}$, and $4^j \le 4^J < 4\varepsilon^{-2}$ since $2^J < 2/\varepsilon$; hence $4^j w(s_j) \le 4\, w(s_j)/\max\{\varepsilon, \Delta(s_j)\}^2 \le 4\,\mathcal{I}_\infty(w;\varepsilon)$. Summing over the $J - j_{\min} + 1$ phases gives the upper bound.

Lower bound. Let $s^\dagger$ attain $\mathcal{I}_\infty(w;\varepsilon)$. Set $j^\dagger := J$ if $\Delta(s^\dagger) \le 2^{-J}$, and $j^\dagger := \lfloor \log_2(1/\Delta(s^\dagger)) \rfloor$ otherwise. In both cases $\Delta(s^\dagger) \le 2^{-j^\dagger}$, and $j_{\min} \le j^\dagger \le J$ because $\Delta(s^\dagger) \le (1-\gamma)^{-1} \le 2^{-j_{\min}}$. If $j^\dagger = J$, then $4^{j^\dagger} \ge \varepsilon^{-2} \ge \max\{\varepsilon, \Delta(s^\dagger)\}^{-2}$. Otherwise $4^{j^\dagger} > \frac{1}{4}\Delta(s^\dagger)^{-2} \ge \frac{1}{4}\max\{\varepsilon, \Delta(s^\dagger)\}^{-2}$. Keeping only the term $j = j^\dagger$ of (4) gives $\mathcal{I}_{\mathrm{ph}}(w;\varepsilon) \ge 4^{j^\dagger} w(s^\dagger) \ge \frac{1}{4}\mathcal{I}_\infty(w;\varepsilon)$.

The logarithmic factor. Take states $s_0, \dots, s_M$ with $\Delta(s_i) = 2^{-i}$ and $w(s_i) = 4^{M-i}$, and $\varepsilon = 2^{-M}$. For $0 \le j \le M$, the inner maximum is $\max_{i \ge j} 4^{M-i} = 4^{M-j}$, so each of these phases contributes 4^M and $\mathcal{I}_{\mathrm{ph}}(w;\varepsilon) \ge (M+1)4^M$. On the other hand, $\mathcal{I}_\infty(w;\varepsilon) = \max_i 4^{M-i} \cdot 4^i = 4^M$. The ratio is at least $M + 1 = \log_2(1/\varepsilon) + 1$. □

M Additional related work

Gap-dependent convergence. Natural policy gradient and policy mirror descent converge linearly when all advantage gaps are bounded away from zero (Xiao, 2022; Li et al., 2024). These results characterize the *optimization speed* of policy updates but do not address the *sample cost of trajectory-based exploration*—in particular, the mixing-time overhead of visiting hard-to-reach states. Our locking argument (Lemma 5.3) instead works through the dual coordinates of the mirror-descent iterates rather than the value error, and its role is to *shrink the active set*.

Instance-dependent RL. In episodic tabular MDPs, gap-dependent *regret* bounds are known (Simchowitz & Jamieson, 2019), and so are instance-dependent *PAC* guarantees, which are closer in spirit to our work: Wagenmaker et al. (2022); Wagenmaker & Jamieson (2022) provide near-optimal instance-dependent rates emphasizing the interplay between reachability and gap dependence; Tirinzoni et al. (2022; 2023) give near-matching bounds for deterministic and general episodic MDPs. Our setting differs: it has no resets, no simulator access, and correlated data. For this setting, Al Marjani et al. (2021) study best-policy identification: they give an instance-dependent lower bound through stationary allocations that respect the navigation constraints, and an algorithm whose sample complexity is instance-dependent, both in the asymptotic regime $\delta \to 0$ and for exact identification of an optimal policy. To our knowledge, no prior work provides non-asymptotic instance-dependent *exploration* bounds for ε-optimal policies in this setting.

Gap-dependent bandits. Our phase-by-phase accounting, in which resolved states leave the active set, mirrors classical gap-dependent bandit analyses (Lai & Robbins, 1985; Auer et al., 2002; Audibert et al., 2010), where arms with large reward gaps are identified early and eliminated from consideration. Our local complexity extends this principle to a Markov chain, where the cost of visiting a state is its certified hitting cost rather than one arm pull.

Lower bounds for single-trajectory RL. Matching information-theoretic lower bounds are classical in the generative-model setting (Gheshlaghi Azar et al., 2013; Agarwal et al., 2020), but single-trajectory lower bounds are harder because the algorithm's policy shapes the data. Our family-specific lower bound is a two-point argument between MDPs whose transition kernels differ at a single state; costs are deterministic, so all information comes from observed transitions, and the chain's visit rate limits how fast the two instances can be distinguished. A general lower bound in terms of hitting costs and gaps remains open for the single-trajectory discounted setting.