Non-Asymptotic Best Policy Identification Guarantees in Online Reinforcement Learning

Joseph Lazzaro
Imperial College London
jl22718@ic.ac.uk

Alessio Russo
Boston University
arusso2@bu.edu

Aldo Pacchiano
Boston University
Broad Institute of MIT and Harvard
pacchian@bu.edu

Abstract

In this work we study the Best Policy Identification (BPI) problem in online, tabular Reinforcement Learning. This is an active sequential hypothesis testing problem in which the learner's objective is to identify an optimal policy in a Markov Decision Process (MDP) with high confidence, while minimizing the expected sample complexity to do so. We consider an online setting with deterministic rewards, where the agent must strategically navigate through the MDP in order to effectively explore. Previous works, including Navigate and Stop (NaS) and its variants, provide instance-dependent asymptotic guarantees for BPI. The finite-confidence behavior of NaS remains less well understood. In this work, we fill that gap by providing the first non-asymptotic sample complexity guarantees for a modified NaS algorithm, showing that its sample complexity depends not only on the characteristic time, but also on the connectivity of the underlying MDP, the sharpness of the allocation objective, and other instance-dependent quantities. We identify these additional attributes and make explicit their contributions to the overall sample complexity.

1 Introduction

In Reinforcement Learning [32], a fundamental question is how efficiently one can identify an optimal control strategy when interacting with an unknown stochastic environment. While most classical work focuses on regret minimization [7, 21, 5], regret is not always the most relevant performance criterion. In many practical scenarios, including robotics prototyping, model selection for planning, and simulation-based decision support, the reward collected during learning is irrelevant; rather, the goal is to *gather information as quickly as possible and stop once a correct decision can be made with high confidence*. This motivates the study of *active pure exploration* in Markov decision processes (MDPs) and, in particular, the problem of *Best Policy Identification* (BPI) [2].

In BPI, the learner interacts with a discounted MDP and sequentially chooses actions and observes transitions, with the reward function known, and aims to identify an optimal policy with probability at least $1-\delta$ while minimizing the number of interaction steps. This problem was formalized in early PAC-RL work and has recently received renewed attention, culminating in the first instance-dependent lower bounds and algorithms for discounted MDPs under navigation constraints [2, 26, 27, 25, 33]. Crucially, unlike the generative model setting, where an algorithm may sample arbitrary state-action pairs at will, the online navigation setting restricts sampling to a single evolving system trajectory, making the structure of optimal exploration policies substantially richer.

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

The paper of Al Marjani et al. [2] investigates BPI under navigation constraints and provides: (i) an instance-specific lower bound on sample complexity, formalized as a game between the algorithm and nature, (ii) a sampling rule targeting the allocation of a tractable oracle, and (iii) an asymptotic sample-complexity analysis in the limit $\delta \to 0$. Their *Navigate-and-Stop* (NaS) algorithm gives an instance-dependent asymptotic guarantee based on a tractable relaxation of the allocation problem, constituting the first such guarantee for online BPI. However, the NaS guarantees motivating this work are *asymptotic*: the theory guarantees optimal scaling as $\delta \to 0$ but provides no understanding of finite-confidence behavior, the effect of stopping thresholds, or the transient instability of sampling distributions when data are scarce.

The need for non-asymptotic analysis has been recently established in pure exploration bandits. In the classical Best-Arm Identification (BAI) problem, Track-and-Stop (TAS) is known to be asymptotically optimal [13]. Yet, despite a decade of work, the finite-δ properties of TAS long remained elusive. Game theoretic approaches with optimistic extensions of TAS were proposed to provide non-asymptotic guarantees [12]. Very recent progress [22] has also illustrated that with careful technical approaches to the analysis of TaS, it is possible to provide non-asymptotic sample complexity guarantees without needing to modify the original algorithm. These developments in BAI raise a natural and important question for BPI:

Can we characterize, and control, the non-asymptotic sample complexity of the Navigate-and-Stop algorithm for MDPs?

Indeed, while bandit pure-exploration methods now admit both asymptotic and finite-confidence guarantees [22, 12], analogous finite-confidence guarantees for NaS remain the focus of this work. The online MDP setting is far more challenging, combining bandit-style change-of-measure arguments with (i) navigation constraints, (ii) non-homogeneous Markov dynamics induced by adaptive policies, and (iii) the need to maintain forced exploration so as to avoid state-space collapse. As a result, controlling finite-time deviations of empirical transition kernels and stopping statistics is substantially more delicate than in bandits.

Contributions. In this paper, we close this gap by providing non-asymptotic sample-complexity guarantees for a NaS-type algorithm in discounted MDPs under navigation constraints:

1. First, we prove the first finite-confidence sample-complexity upper bound for a NaS-type algorithm in single-trajectory discounted tabular MDPs. Under our assumptions, the algorithm is δ-correct and satisfies an implicit non-asymptotic bound of the form

$$\mathbb{E}[\tau] \leq \inf\{n \geq t_M : \ \forall t \geq n, \ \overline{b}(t, \delta) < tT(M)^{-1} - \mathcal{D}_M(t)\} + O(1),$$

where $\mathcal{D}_M(t) = o(t)$ is an error term driven by the finite-confidence costs of estimating the transition probabilities, navigating the MDP under a single trajectory, and tracking the optimal state-action allocation induced by the estimated model.

2. Under linear sharpness of the allocation objective, we obtain an explicit finite-confidence bound with leading term $T(M)\log(1/\delta)$, a confidence-independent initial-period cost, and sublinear corrections with powers $1-(1-\alpha)^m/4$ and $1-\alpha$, where m is one plus the support-graph diameter. Their coefficients expose the dependence on transition coverage, target-chain behavior, forced exploration, and allocation geometry; Corollary 10 specifies the complete bound at every confidence level.

3. Third, we develop the technical ingredients needed for this result: finite-time tracking of empirical visitation frequencies under adaptive non-homogeneous Markov dynamics, stability of optimal-allocation selectors under transition perturbations, and a sharpness analysis showing that the objective can be arbitrarily flat in continuum alternative classes.

2 Problem Setting

In this section, we introduce the problem setting, concepts and notation used throughout the paper.

Markov Decision Processes (MDPs). We consider discounted MDPs denoted by $M = (\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ is the finite state space (of size $S = |\mathcal{S}|$), $\mathcal{A}$ is the finite action space (of size $A = |\mathcal{A}|$), $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition function, which maps state-action pairs to a distribution over states, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ defines a mapping from state-action pairs to deterministic real rewards, and $\gamma \in (0, 1)$ is the discount factor. For a reward r, we define the discounted

value of a Markov policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ to be $V^\pi(s) = \mathbb{E}^\pi[\sum_{t\geq 0} \gamma^t r(s_t, a_t)|s_0 = s]$, where $s_{t+1} \sim P(\cdot|s_t, a_t)$ and $a_t \sim \pi(\cdot|s_t)$. We also let $V^\star(s) = \max_\pi V^\pi(s)$ be the optimal value in s and, for MDPs with a unique optimal policy, we denote it by $\pi_o^\star$. Similarly, we also define the action-value function $Q^\pi(s, a) = r(s, a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)}[V^\pi(s')]$. For an MDP M, we indicate by $\Pi^\star(M) = \{\pi \in \Pi_d : \|V^\star - V^\pi\|_\infty = 0\}$ the set of optimal policies in M and Π_d is the set of deterministic policies.

Best Policy Identification with Fixed Confidence Best Policy Identification consists of finding an optimal policy in an MDP using the least number of samples with a given confidence level for a specific reward function [1, 2]. The BPI objective can be formalized in the δ-Probably-Correct (δ-PC) framework, where the learner interacts with the MDP until they can output an optimal policy with confidence $1 - \delta$. Formally, an algorithm in our δ-PC setting consists of: (i) an exploration strategy (a.k.a. sampling rule), (ii) a stopping rule $\tau = \tau_\delta$, whose subscript records the confidence level, and (iii) an estimated optimal policy $\hat{\pi}_\tau$. At each time step t, the algorithm observes a state s_t, it selects an action a_t, and then observes the next state $s_{t+1} \sim P(\cdot|s_t, a_t)$. We denote by $\mathbb{P}$ (resp. $\mathbb{E}$) the probability law (resp. expectation) of the process $(X_t)_t$, where $X_t = (s_t, a_t)$. We use the pre-action history $\mathcal{F}_t = \sigma(s_0, a_0, s_1, \ldots, a_{t-1}, s_t)$; the action probabilities at time t are measurable with respect to this history. For such algorithm, we define τ to be a stopping rule w.r.t. the filtration $(\mathcal{F}_t)_{t\geq 1}$. This stopping rule τ simply states when to stop the exploration phase. At the stopping time τ the algorithm outputs an estimate $\hat{\pi}_\tau$ of the optimal policy using the collected data. We focus on δ-PC algorithms, according to the following definition.

Definition 1 (δ-PC Algorithm)**.** *We say that an algorithm is δ-PC if, for every M in the specified admissible class $\mathcal{M}$, it outputs (in finite time) an optimal policy with probability at-least $1 - \delta$, i.e., $\mathbb{P}[\tau < \infty,\ \hat{\pi}_\tau \in \Pi^\star(M)] \geq 1 - \delta$.*

Let $N_t(s, a)$ count the visits to (s, a) among the t completed transitions, before selecting action t. The expected sample complexity of an algorithm $\mathcal{A}$ can be written as $\mathbb{E}[\tau] = \sum_{s,a} \mathbb{E}[N_\tau(s, a)]$. It is this sample complexity that the learner hopes to minimize.

2.1 Sample Complexity Lower Bound

For BPI, the following lower bound Theorem 2 has been shown to hold for all δ-PC algorithms in [2], where they also provided asymptotic guarantees on their algorithm NaS. Before stating the bound, we first recall two structural constraints on the learning problem. These are not assumptions imposed by our analysis, but rather necessary conditions inherited from the information-theoretic lower bound that any δ-PC algorithm must satisfy. The *information constraints* arise from change-of-measure arguments: to guarantee correctness, the algorithm must collect enough statistical evidence to distinguish the true MDP from every alternative with a different optimal policy. The *navigation constraints* are specific to the online setting: since observations are generated along a single trajectory, the expected visitation counts must satisfy the flow conservation conditions of the induced Markov chain. Together, these two constraints characterize the feasible set of sampling allocations for any δ-PC algorithm.

(i) Information constraints. The algorithm must collect enough evidence to distinguish the true MDP M from any *alternative* MDP M' with a different optimal policy. Define the alternative set

$$\mathrm{Alt}(M) = \{M' \in \mathcal{M} : M \ll M', \Pi^\star(M) \cap \Pi^\star(M') = \emptyset\}, \tag{1}$$

where $M \ll M'$ means that every transition row of M is absolutely continuous with respect to the corresponding row of M'. Then, for $0 < \delta < 1/2$, any δ-PC algorithm, and any $M' \in \mathrm{Alt}(M)$,

$$\sum_{(s,a)} \mathbb{E}_M[N_{\tau_\delta}(s, a)]\, \mathrm{KL}_{M|M'}(s, a) \geq \mathrm{kl}(\delta, 1 - \delta),$$

where we define these KL terms to be the KL divergence between the two respective transition functions at each state action pair, $\mathrm{KL}_{M|M'}(s, a) = \mathrm{KL}(P_M(\cdot|s, a)||P_{M'}(\cdot|s, a))$ and kl is the binary divergence. We also define the KL divergence between two transitions as the sum of the KL divergences of the transitions at each state-action pair $\mathrm{KL}(M, M') = \sum_{s,a} \mathrm{KL}_{M|M'}(s, a)$.

(ii) Navigation constraints. Since observations arise from a single controlled trajectory, expected visitation counts must satisfy stationarity conditions of the induced Markov chain. Specifically,

whenever $\mathbb{E}_M[\tau_\delta] < \infty$, for all $s \in \mathcal{S}$,

$$\left| \mathbb{E}_M[N_{\tau_\delta}(s)] - \sum_{s',a'} P_M(s|s',a') \, \mathbb{E}_M[N_{\tau_\delta}(s',a')] \right| \leq 1.$$

As $\delta \to 0$ and the expected stopping time diverges, every subsequential limit of the normalized expected counts satisfies exact flow conservation.

We can now state the asymptotic lower bound on the sample complexity of any δ-PC algorithm for the BPI problem. This is the navigation-constrained lower bound of Al Marjani et al. [2], specialized to known rewards and the same admissible class used by our algorithm; Appendix C.7 records this specialization. We begin by defining the set of *navigation-constrained allocations*:

$$\Omega(M) = \left\{ \omega \in \Delta(S \times A) : \omega(s) = \sum_{s',a'} P_M(s|s',a') \omega(s',a'), \ \forall s \in \mathcal{S} \right\},$$

where $\omega(s) = \sum_a \omega(s,a)$. We can then state the lower bound as follows.

Theorem 2 (Asymptotic lower bound; 2, Proposition 2)**.** *For any algorithm $\mathcal{A}$ that is δ-PC on $\mathcal{M}$ and any $M \in \mathcal{M}$ with nonempty alternatives,*

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_M[\tau_\delta]}{\log(1/\delta)} \geq T(M),$$

where

$$\frac{1}{T(M)} = \sup_{\omega \in \Omega(M)} \inf_{M' \in \mathrm{Alt}(M)} \sum_{(s,a)} \omega(s,a) \, \mathrm{KL}_{M|M'}(s,a).$$

This lower bound can be interpreted as a two-player zero-sum game: the learner selects an allocation $\omega \in \Omega(M)$, while nature chooses the hardest alternative M'. Compared to the classical bandit setting, the key distinction is that $\Omega(M)$ enforces feasibility via the MDP dynamics, restricting allocations to stationary distributions. Consequently, the achievable rate is no better, and typically strictly worse, than in the generative model setting, where arbitrary sampling is allowed.

For convenience, define

$$T(\omega, M)^{-1} = \inf_{M' \in \mathrm{Alt}(M)} \sum_{(s,a)} \omega(s,a) \, \mathrm{KL}_{M|M'}(s,a).$$

Write $J_M(\omega) = T(\omega, M)^{-1}$ and $J_M^* = \max_{\omega \in \Omega(M)} J_M(\omega) = 1/T(M)$. Let $\Omega^\star(M) = \arg\max_{\omega \in \Omega(M)} J_M(\omega)$ denote the set of optimal allocations, and define a selector

$$\omega^*(M) = \operatorname*{arg\,min}_{\omega \in \Omega^\star(M)} \|\omega\|_2^2.$$

We omit dependence on M when clear. Note that $\Omega^*(M)$ is a convex set (this is proven in Lemma 28) and thus $\omega^*(M)$ is well defined.

Remark 3 (Non-uniqueness)**.** *The optimal allocation set $\Omega^\star(M)$ need not be a singleton. Prior work often avoids this issue through convex relaxations or approximations, which can yield a unique target allocation but may sacrifice optimality [e.g. 2, 27, 25]. We instead work directly with the navigation-constrained allocation problem, allowing $\Omega^\star(M)$ to be set-valued. This affects both the analysis of NaS and the sampling rule used to track an optimal allocation. Non-singleton examples are discussed in Theorem 4.*

2.2 Assumptions

In this work, we study the behavior of the sample complexity $\mathbb{E}_M[\tau_\delta]$ and, in particular, provide non-asymptotic sample-complexity guarantees for a modified version of Navigate and Stop [2]. To do so, we impose the following assumptions and discuss them below. Throughout, the reward function and discount factor are fixed and known, $S, A \geq 2$, and $\mathcal{M}$ is the known admissible model class used in both the lower bound and the algorithm. The cases $S = 1$ or $A = 1$ are trivial with known rewards and can be solved without sampling.

Assumption 1. *(I) $M \in \mathcal{M}$, where $\mathcal{M}$ is a set of communicating MDPs with a unique optimal deterministic policy; (II) M is aperiodic under a policy assigning positive probability to every action in every state.*

Assumption 2. *There exist a known reference transition kernel μ and a known $L > 1$ such that every admissible model has its transition kernel in*

$$G_\mu(L) = \left\{ Q \in \Delta(S)^{S \times A} : \ \mu(s'|s,a)/L \le Q(s'|s,a) \le L\mu(s'|s,a) \ \ \forall s, a, s' \right\}. \tag{2}$$

In particular, these bounds apply to all alternatives in $\mathrm{Alt}(M)$.

Assumption 3 (Irreducibility of the selected target)**.** *The set* $\mathrm{Alt}(M)$ *is nonempty. Let* $x = \omega^*(M)$ *be the minimum-norm optimal allocation and define* $\pi^\dagger(a|s) = x(s,a)/x(s)$ *when* $x(s) > 0$*, using a uniform row otherwise. The state kernel*

$$K^\dagger(s,s') = \sum_a \pi^\dagger(a|s) P_M(s'|s,a)$$

is irreducible.

Assumption 1 is standard in the BPI literature [2, 26, 27, 25]. Communication ensures that exploration can reach every state. Aperiodicity is assumed as in the usual setting, although the fixed-Poisson-equation argument below only requires irreducibility of the target chain. The unique optimal policy requirement also admits a generic justification. Under a non-degeneracy condition on the known rewards, specified in Appendix E, we prove that the set of transition kernels admitting multiple optimal policies has Lebesgue measure zero in the product of transition simplices. Consequently, uniqueness holds almost surely under any prior absolutely continuous with respect to this measure. This result may be of independent interest, particularly for Bayesian sequential testing problems.

Assumption 2 is analogous to boundedness assumptions in bandit best-arm identification [e.g. 12, 22]. It provides the learner with knowledge of transition support and bounds the density ratios on that support. These bounds ensure that the KL divergences and the allocation objective are well behaved. The set $G_\mu(L)$ is convex; see Lemma 22. The class $\mathcal{M}$ may be a structured subset of $G_\mu(L)$, but the same class must be used throughout. Thus $T(M)$ is the characteristic time relative to the learner's known class, rather than to unrestricted alternatives.

Finally, Assumption 3 addresses the case where the set of policy-changing alternatives obtained by modifying a particular state-action pair may be empty. This can lead to optimal visitation strategies assigning no mass to some pairs. Allowing policy-changing reward perturbations in the more general stochastic-reward setting, as in Al Marjani et al. [2], can avoid this obstacle. With known rewards, we require that the selected exploration strategy can still reach every state, while allowing zero mass on some actions. Appendix C.10.1 proves sufficient conditions using admissible local alternatives. In particular, full transition support guarantees the irreducibility part of Assumption 3; nonempty alternatives remain a separate requirement. The assumption lets us focus on non-asymptotic guarantees depending on the underlying dynamics. Under Assumptions 1-3, $0 < J_M^* < \infty$.

3 Navigate-and-Stop Algorithm

In this section we describe the Navigate-and-Stop (NaS) algorithm [2] for which we provide non-asymptotic analysis. The algorithm is motivated by the instance-specific lower bound on sample complexity in Section 2.1. Since the optimal visitation frequency depends on the unknown model M, NaS follows an estimate-then-track strategy: at each round, it estimates the transition kernel from data, computes an allocation for the current estimate, and plays according to the induced navigating policy. To ensure sufficient coverage for consistent estimation, this policy is mixed with a forcing component. Our variant handles non-uniqueness through a relaxed minimum-norm selector and uses a state-dependent forcing schedule. The relaxation stabilizes a single limiting allocation; the fixed-Poisson analysis then controls its visitation error through cumulative policy errors. This direct tracking argument requires irreducibility of the selected target. Thus our result concerns this specified variant of NaS. The stopping rule uses the raw empirical likelihood, while the allocation oracle uses the projected estimate.

Algorithm 1 NaS with a relaxed allocation selector

Require: Confidence $\delta \in (0,1)$; $\alpha \in (0,1)$, $\beta \geq 0$, $\nu \in (0,1)$; $\mu, L, \mathcal{M}$.
Initialize $t = 0$, all counts to zero, and observe s_0.
while true **do**
  Form $\widehat{P}_t$, its projection P_t^{Π}, and a deterministic optimal policy $\widehat{\pi}_t^r$ of M_t.
  if $t \geq 1$ and $Z_t \geq b(N_t, \delta)$, with Z_t from (5) **then**
    return $\widehat{\pi}_t^r$.
  end if
  Compute r_t, η_t and $\omega_t^\star$ by (4), with the stated fallback if $\mathcal{A}(\widehat{\pi}_t^r) = \varnothing$.
  Set $\varepsilon_t(s_t) = \max\{1+\nu, N_t(s_t)\}^{-\alpha}$ and
    $\pi_t(\cdot|s_t) = (1-\varepsilon_t(s_t))\pi_t^{\text{base}}(\cdot|s_t) + \varepsilon_t(s_t)\pi_{f,t}(\cdot|s_t)$.
  Draw $a_t \sim \pi_t(\cdot|s_t)$ and observe $s_{t+1} \sim P_M(\cdot|s_t, a_t)$.
  Update the counts and set $t \leftarrow t+1$.
end while

3.1 Algorithm

NaS is made of three components: (1) a sampling rule that dictates how to choose the next action; (2) a stopping rule that dictates when to stop collecting data; and (3) a recommendation rule that outputs an estimated optimal policy. We describe these components below and summarize them in Algorithm 1. Counts are taken before action t, so $N_t(s,a) = \sum_{j=0}^{t-1} \mathbb{1}\{(s_j, a_j) = (s,a)\}$ and $\sum_{s,a} N_t(s,a) = t$.

Empirical estimates of the MDP. Define $N_t(s,a,s') = \sum_{j=0}^{t-1} \mathbb{1}\{(s_j, a_j, s_{j+1}) = (s,a,s')\}$ and $N_t(s,a) = \sum_{s'} N_t(s,a,s')$. The raw empirical kernel is $\widehat{P}_t(s'|s,a) = N_t(s,a,s')/N_t(s,a)$ for a visited row, and $\mu(s'|s,a)$ otherwise. We project it onto the known bounded class:

$$P_t^{\Pi} \in \operatorname*{arg\,min}_{Q \in G_\mu(L)} \sum_{s,a} \|\widehat{P}_t(\cdot|s,a) - Q(\cdot|s,a)\|_1. \tag{3}$$

Choose a fixed measurable tie-breaking rule, for example minimum squared Euclidean norm among the minimizers. Since $G_\mu(L)$ is a product of row sets, the projection minimizes each row distance. Let $M_t = (\mathcal{S}, \mathcal{A}, P_t^{\Pi}, r, \gamma)$ and choose $\widehat{\pi}_t^r$ by a fixed tie-breaking rule among its deterministic optimal policies. The estimate need not itself belong to a structured $\mathcal{M}$.

Sampling rule. For any deterministic policy π, write $\mathcal{A}(\pi) = \{N \in \mathcal{M} : \pi \notin \Pi^*(N)\}$ for the set of admissible models under which recommending π would be incorrect. For an estimated model N, let $\pi_o(N)$ denote an optimal deterministic policy selected using the fixed tie-breaking rule, so that $\widehat{\pi}_t^r = \pi_o(M_t)$. Extend the information objective to the whole allocation simplex by

$$J_N(\omega) = \inf_{N' \in \mathcal{A}(\pi_o(N))} \sum_{s,a} \omega(s,a) \mathrm{KL}(P_N(\cdot|s,a), P_{N'}(\cdot|s,a)).$$

This agrees with $T(\omega, N)^{-1}$ for $N \in \mathcal{M}$. Define $J_N^* = \max_{\omega \in \Omega(N)} J_N(\omega)$. Instead of exact tie-breaking over its maximizers, use

$$\Omega_{\eta_t}(M_t) = \{\omega \in \Omega(M_t) : J_{M_t}(\omega) \geq J_{M_t}^* - \eta_t\}, \qquad \omega_t^\star = \operatorname*{arg\,min}_{\omega \in \Omega_{\eta_t}(M_t)} \tfrac{1}{2}\|\omega\|_2^2. \tag{4}$$

The set is compact and convex, so the selector is unique. Put $n_t = \min_{s,a} N_t(s,a)$ and, for $t \geq 1$,

$$r_t = \begin{cases} 2, & n_t = 0, \\ \min\left\{2, \sqrt{\dfrac{8}{n_t} \log\big(2^S SA\, t(t+1)^7\big)}\right\}, & n_t > 0, \end{cases} \qquad \eta_t = \log(e+t) r_t,$$

with $r_0 = 2$ and $\eta_0 = 2$. This tolerance is computable without unknown instance constants and tends to zero almost surely. Set $\pi_t^{\text{base}}(a|s) = \omega_t^\star(s,a) / \sum_b \omega_t^\star(s,b)$, using a uniform row if the denominator vanishes. If $\mathcal{A}(\widehat{\pi}_t^r) = \varnothing$ during initialization, use a uniform base policy.

As in Russo and Vannella [27], the forcing policy encourages selecting under-sampled actions:

$$\pi_{f,t}(\cdot|s) = \mathrm{softmax}(-\beta_t(s) N_t(s,\cdot)), \qquad \text{where}$$

$$\beta_t(s) = \begin{cases} \dfrac{\beta \log(\max\{1, N_t(s)\})}{\max_a N_t(s,a) - \min_a N_t(s,a)}, & \max_a N_t(s,a) > \min_a N_t(s,a), \\ 0, & \text{otherwise.} \end{cases}$$

Every minimum-count action receives forcing probability at least $1/A$. The mixture weight $\varepsilon_t(s) = \max\{1+\nu, N_t(s)\}^{-\alpha}$ depends only on the local state count. The coverage proof below works for every $0 < \alpha < 1$ and $\beta \geq 0$; it does not require $\alpha + \beta < 1$.

Stopping and recommendation rules. The stopping statistic compares the raw empirical likelihood with models on which the proposed recommendation is incorrect:

$$Z_t = \inf_{N \in \mathcal{A}(\widehat{\pi}_t^r)} \sum_{s,a} N_t(s,a) \mathrm{KL}(\widehat{P}_t(\cdot|s,a), P_N(\cdot|s,a)). \tag{5}$$

An empty alternative set gives $Z_t = +\infty$, in which case the recommendation is valid throughout $\mathcal{M}$; an unvisited row contributes zero. With

$$b(N_t, \delta) = \log(1/\delta) + (S-1) \sum_{s,a} \log\left(e \left[1 + \frac{N_t(s,a)}{S-1} \right] \right),$$

stop at $\tau_\delta = \inf\{t \geq 1 : Z_t \geq b(N_t, \delta)\}$ and return $\widehat{\pi}^r_{\tau_\delta}$. Error control follows from the standard transition-likelihood mixture argument [2, Lemma 15]. Lemma 19 gives the proof for this statistic and separately establishes finite stopping under our sampling rule.

4 Non-Asymptotic Upper Bound for NaS

4.1 Sharpness of the Objective Function

Before presenting formal sample-complexity bounds for Algorithm 1, we isolate a key finite-confidence cost: the rate at which the empirical allocations induced by the estimated oracle policies π_t^{base} approach an optimal allocation. This rate depends on the local geometry of $T(\omega, M)^{-1}$ near its maximizers. When the objective is flat, small model-estimation errors can induce large allocation errors, delaying the stopping rule. We capture this effect through a *sharpness function*.

Sharpness Function. Recall $J_M(\omega) = T(\omega, M)^{-1}$, its maximum J_M^*, and its optimal set $\Omega^\star(M)$. Write $\mathrm{dist}_\infty(\omega, \Omega^\star(M)) = \inf_{\omega' \in \Omega^\star(M)} \|\omega - \omega'\|_\infty$. Let φ be nondecreasing with $\varphi(0) = 0$ and $\lim_{u \downarrow 0} \varphi(u) = 0$. Then φ is a sharpness function under M if $\varphi(J_M^\star - J_M(\omega)) \geq \mathrm{dist}_\infty(\omega, \Omega^\star(M))$ for every $\omega \in \Omega(M)$. We say that J_M has *polynomial sharpness of order* $p > 0$ if there exists a constant $c > 0$ such that $\varphi(t) \leq c\, t^p$. A local polynomial bound can be extended to a global modulus by increasing its constant. The selector analysis uses a sharpness modulus for the true MDP M, without requiring a common modulus valid uniformly over nearby models.

Theorem 4 shows that this effect can be arbitrarily severe: for any $m > n \geq 1$, there is a structured admissible family with local polynomial sharpness exponent $p = (m-n)/m$. Figure 1 verifies these rates numerically; details are deferred to Appendix D. We visualize the simple family of MDPs used in $\mathcal{M}$ in Figure 2 and provide more details of the theorem's proof and construction in Appendix D.

Theorem 4. *Fix integers $m > n \geq 1$. There exist an MDP M in an admissible family $\mathcal{M}$ and a constant $c_{m,n} > 0$ such that, for every sufficiently small $d > 0$, there is an allocation $\omega^{(d)} \in \Omega(M) \setminus \Omega^\star(M)$ with $\mathrm{dist}_\infty(\omega^{(d)}, \Omega^\star(M)) = d$ and*

$$J_M^\star - J_M(\omega^{(d)}) = c_{m,n} d^{\frac{m}{m-n}} + o\big(d^{\frac{m}{m-n}}\big), \qquad d \downarrow 0.$$

Moreover, $\varphi_M(u) = C u^{(m-n)/m}$ is a valid sharpness modulus for some $C > 0$, and no polynomial modulus with a larger exponent is valid.

It may be of practical relevance to restrict the set of MDPs to be within some discrete set, for example due to some limitations of precision. In this case, we can provide some simplifications, showing that the sharpness becomes linear. The proof can be found in Appendix D.3.

Proposition 5. *If the admissible class $\mathcal{M}$ is finite and $\mathrm{Alt}(M)$ is nonempty, a linear sharpness modulus exists.*

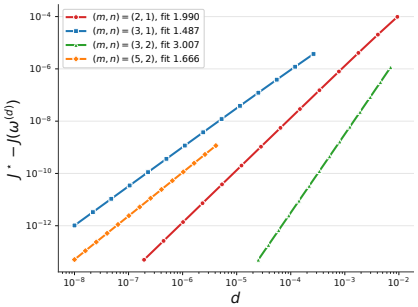

Figure 1: Empirical sharpness scaling for the MDP family of Theorem 4. For each choice of integers $m > n \geq 1$, we plot the objective gap $J^\star - J(\omega^{(d)})$ with different values of d. On log-log axes, the slope of each curve estimates the gap exponent $m/(m-n)$, and hence the reciprocal slope estimates the sharpness exponent of the function φ, namely $(m-n)/m$. The different choices of (m, n) therefore realize different polynomial sharpness exponents.

We note that the definition of sharpness in this section, which appears later in the sample complexity upper bound is representative of the worst case sharpness of $J_M(\omega)$ approaching the set of optimal allocations from any direction. However, it would be interesting future work to see if there was an algorithmic formulation which could exploit more approachable paths to the optimal set $\Omega^*(M)$ for improved sample complexity.

4.2 Implicit and Explicit Upper Bounds

We now give non-asymptotic bounds for Algorithm 1, separating estimation, navigation, and allocation costs. Under linear sharpness, the implicit bound yields an explicit finite-confidence bound.

Write $\rho_M = \min_s \omega^*(M)(s)$ for the smallest target state frequency. The constant $\kappa_\Sigma = \kappa_0 + \kappa_1$ measures how feasible exploration allocations change under transition perturbations (Lemma 31). The constant H_M controls transient visitation errors under the target exploration policy and the accumulation of policy errors; its formal definition is the maximum Poisson span in (23). These are fixed instance-dependent quantities, independent of time and confidence. Define

$$K_M = 2SA \log L\,(1 + H_M), \qquad C_M = 2 + 4 \log L + 2SA \log L\,\kappa_\Sigma.$$

Let m be one plus the maximum shortest-path length in the state support graph and put $\zeta = (1-\alpha)^m$. Define the transition-estimation envelope

$$e_t = C_\psi t^{-\zeta/2} \sqrt{\log(et)}, \qquad t \geq 1.$$

Here C_ψ, defined in (34), measures the difficulty of sampling every transition row and increases as coverage weakens. It depends on the transition dynamics and α, but not on t or δ. Choose an integer t_M satisfying (42), after which coverage, policy identification, and allocation stability can be controlled together. This initial time is finite and independent of δ, but depends on the instance, algorithm parameters, and sharpness modulus. For $t \geq t_M$, e_t bounds transition-estimation error with high probability.

Theorem 6. *Under Assumptions 1-3, for every $\delta \in (0,1)$,*

$$\mathbb{E}_M[\tau_\delta] \leq \inf\left\{ n \in \mathbb{N} : n \geq t_M,\ \forall t \in \mathbb{N},\ t \geq n :\ \bar{b}(t,\delta) < \frac{t}{T(M)} - \mathcal{D}_M(t) \right\} + 1, \tag{6}$$

where $\bar{b}(t,\delta) = \log(1/\delta) + c_1 \log t + c_2$, $c_1 = SA(S-1)$, $c_2 = c_1(1 + \log 2)$, and

$$\begin{aligned} \mathcal{D}_M(t) = (1 + \log(2L^2))\, te_t \\ + K_M \Bigg[t_M + S + 2 + d_{S,A}\sqrt{t \log(et)} + \frac{S^\alpha}{1-\alpha} t^{1-\alpha} \\ + \frac{A(A+1)}{2\rho_M} \sum_{j=1}^{t-1} \sqrt{\kappa_\Sigma e_j + \varphi_M\big((\log(e+j) + C_M)e_j\big)} \Bigg], \end{aligned} \tag{7}$$

with $d_{S,A} = 1 + \sqrt{(\log(2SA) + 3)/2}$. In particular, $\mathcal{D}_M(t) = o(t)$.

The first term in $\mathcal{D}_M$ is the current transition-estimation cost. The bracket describes the difficulty of realizing the target allocation along a trajectory: the initial period, martingale fluctuations, forced exploration, and accumulated allocation error. A larger sharpness modulus weakens allocation stability, increasing the last sum and potentially the initial period. The factor $1/\rho_M$ controls normalization into action probabilities without requiring every action to have positive target mass. Connectivity affects transition coverage through ζ and C_ψ, while H_M captures the behavior of the selected target chain. The additive 1 bounds the universal failure term $C_{\mathrm{tail}} \approx 0.60274 < 1$ in (30). The forcing temperature β does not enter these bounds, since minimum-count actions receive forcing probability at least $1/A$.

Remark 7 (Stochastic rewards)**.** *Extending the analysis to unknown stochastic rewards would require reward-estimation terms in the first component of $\mathcal{D}_M(t)$ and in the perturbation entering φ_M. Under suitable concentration and regularity assumptions on the reward family, these terms could be controlled similarly; variance-sensitive concentration would introduce dependence on reward variances. The characteristic time and stopping threshold would correspond to the joint reward-and-transition model. The navigation argument would have the same form for an irreducible target, although the target allocation and its associated constants could change. We focus on known rewards to isolate the cost of learning and navigating the transition dynamics.*

Since $\mathcal{D}_M(t) = o(t)$, the bound recovers the optimal leading coefficient of $\log(1/\delta)$. Lemma 19 establishes error control and finite stopping on the same admissible class as the lower bound.

Theorem 8 (Asymptotic sample complexity)**.** *Algorithm 1 is δ-PC on the class in Assumptions 1-2. At every instance additionally satisfying Assumption 3,*

$$\lim_{\delta \downarrow 0} \frac{\mathbb{E}_M[\tau_\delta]}{\log(1/\delta)} = T(M).$$

Remark 9 (Comparison with the original NaS guarantee)**.** *The sublinear error in Theorem 6 yields the exact leading coefficient $T(M)$, matching the lower bound for our admissible class. We optimize the information objective without a relaxation of the allocation problem, and the transition-only likelihood test has threshold coefficient one. In comparison, the original NaS guarantee has coefficient $2U_o(M)$, reflecting a tractable allocation relaxation and separate reward and transition thresholds [2, Theorem 6 and footnote 6]. A joint test can avoid the separate-threshold loss with unknown rewards, as discussed there; the characteristic times also refer to their respective known- or unknown-reward classes.*

Under linear sharpness, the sum in (7) has a polynomial bound. Its power also dominates the current-estimation and martingale powers, so we can combine those contributions into a single statistical correction. This slightly looser form leaves two time-dependent costs visible: learning a stable informative allocation and forcing sufficient exploration.

Corollary 10 (Linear sharpness)**.** *Assume the conditions of Theorem 6 and $\varphi_M(u) \le C_\varphi u$ for all $u \ge 0$. Put $a = \zeta/4$ and*

$$K_\varphi = \sqrt{C_\psi[\kappa_\Sigma + C_\varphi(2 + C_M)]}, \qquad C_{\mathrm{stat}} = (1+\log(2L^2))C_\psi + K_M\left[d_{S,A} + \frac{A(A+1)K_\varphi}{2\rho_M(1-a)}\right],$$

$$C_{\mathrm{force}} = \frac{K_M S^\alpha}{1-\alpha}, \qquad B_\delta = \max\{1, T(M)[\log(1/\delta) + c_2]\}.$$

With the confidence-independent initial-period cost C_{init} specified in (43)*, for every $\delta \in (0,1)$,*

$$\begin{aligned}\mathbb{E}_M[\tau_\delta] \le B_\delta + C_{\mathrm{init}} &+ T(M)c_1 \log(2eB_\delta) \\ &+ T(M)C_{\mathrm{stat}}(2B_\delta)^{1-\zeta/4}[\log(2eB_\delta)]^{3/4} \\ &+ T(M)C_{\mathrm{force}}(2B_\delta)^{1-\alpha}.\end{aligned} \tag{8}$$

The bound separates the leading confidence cost, the initial period, and two sublinear corrections. The coefficient C_{stat} displays the effects of allocation sensitivity and target-chain behavior, with C_φ entering through a square root. The powers $1-\zeta/4$ and $1-\alpha$ reveal the influence of coverage and forcing. The term C_{init} accounts for the initial estimation period and the scale needed to solve the implicit stopping inequality. It depends on the problem instance and algorithm parameters, including coverage and sharpness, but is independent of δ. It therefore does not affect the asymptotic leading coefficient, the complete construction in (43) also specifies the initial estimation and logarithmic factors. The proofs are given in Appendix C.9. A sharper decomposition and an extension to polynomial sharpness are given in Appendix C.8.

5 Related Work & Discussion

Many ideas in BPI are inherited from best-arm identification and sequential testing in bandits, most notably Track-and-Stop [13], which combines an allocation-tracking sampling rule with a likelihood-based stopping rule. This perspective has since been extended to settings with multiple correct answers [e.g. 11, 23], alternative objectives [e.g. 20, 18, 34, 17, 14], and richer feedback structures [e.g. 28]. More recently, non-asymptotic analyses have been developed for game-theoretic variants of Track-and-Stop [12] and for the original Track-and-Stop algorithm itself [22], with techniques that motivate our analysis. In RL, much of the classical literature focuses on regret minimization [e.g. 5, 21, 7], whereas BPI is a fixed-confidence sequential testing problem whose goal is to identify an optimal policy as quickly as possible. Existing BPI methods include generative-model approaches [e.g. 1, 33] and online single-trajectory methods, most prominently Navigate-and-Stop [2], with extensions to multi-reward settings [27, 25]. However, in the online NaS setting, existing guarantees are asymptotic (see Appendix A for a more detailed overview).

A main takeaway of this work is that asymptotic optimality alone does not fully characterize the difficulty of online BPI. Even when an algorithm attains the optimal leading coefficient $T(M)$, its finite-confidence performance can be governed by additional instance-dependent effects, including the cost of transition coverage, the connectivity and mixing properties of the MDP, and the sharpness of the allocation objective. Our bounds make these effects explicit, suggesting concrete directions for improved exploration strategies that preserve asymptotic optimality while reducing finite-time costs.

A natural next step is to understand whether the dependencies in our non-asymptotic upper bound can be improved, ideally while preserving the asymptotic optimality of the algorithm. More broadly, we believe that our bounds help identify the aspects of problem instances that must be controlled in the development of new BPI methods. To complement these upper bounds, it would also be interesting to study tight, non-asymptotic lower bounds for BPI. Concurrent work has considered such lower bounds for related problems in sequential testing, but significant work would be required to obtain tight, expressive, non-asymptotic lower bounds in our setting [Appendix A.4, 30]. Another important direction for future work is the development of methods that are not only asymptotically optimal, but also computationally efficient and practically deployable. This is a current limitation of NaS-style strategies. A further promising avenue is to extend these non-asymptotic guarantees to the linear reinforcement learning setting, where richer function approximation classes are considered. We conclude by noting that this work fills a fundamental gap in the BPI literature by establishing finite-confidence guarantees for instance-optimal methods, and we hope it serves as a foundation for further advances in both theory and practice.

Acknowledgments and Disclosure of Funding

Joseph Lazzaro was supported by the Roth Scholarship through the Department of Mathematics at Imperial College London.

References

[1] A. Al Marjani and A. Proutiere. Adaptive sampling for best policy identification in markov decision processes. In *International Conference on Machine Learning*, pages 7459–7468. PMLR, 2021.

[2] A. Al Marjani, A. Garivier, and A. Proutiere. Navigating to the best policy in markov decision processes. In *Advances in Neural Information Processing Systems*, volume 34, pages 25852–25864, 2021.

[3] K. Ariu, P.-A. Wang, A. Proutiere, and K. Abe. Policy testing in markov decision processes. *arXiv preprint arXiv:2505.15342*, 2025.

[4] C. Berge. *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces, and Convexity*. Courier Corporation, 1997.

[5] V. Boone and O.-A. Maillard. The regret lower bound for communicating Markov Decision Processes. In *arXiv preprint arXiv:2501.13013*, 2025.

[6] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities : a non asymptotic theory of independence*. Oxford University Press, 2013.

[7] A. N. Burnetas and M. N. Katehakis. Optimal adaptive policies for markov decision processes. *Mathematics of Operations Research*, 22:222–255, 1997.

[8] G. E. Cho and C. D. Meyer. Comparison of perturbation bounds for the stationary distribution of a Markov chain. *Linear Algebra and its Applications*, 335(1):137–150, 2001.

[9] T. M. Cover and J. A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0471241954.

[10] C. Dann, T. Lattimore, and E. Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. *Advances in Neural Information Processing Systems*, 30, 2017.

[11] R. Degenne and W. M. Koolen. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.

[12] R. Degenne, W. M. Koolen, and P. Ménard. Non-asymptotic pure exploration by solving games. *Advances in Neural Information Processing Systems*, 32, 2019.

[13] A. Garivier and E. Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pages 998–1027. PMLR, 2016.

[14] A. Garivier, P. Ménard, L. Rossi, and P. Menard. Thresholding bandit for dose-ranging: The impact of monotonicity. *arXiv preprint arXiv:1711.04454*, 2017.

[15] A. J. Hoffman. On approximate solutions of systems of linear inequalities. *Journal of research of the National Bureau of Standards*, 49:263, 1952.

[16] A. Jonsson, E. Kaufmann, P. Ménard, O. Darwiche Domingues, E. Leurent, and M. Valko. Planning in markov decision processes with gap-dependent sample complexity. *Advances in Neural Information Processing Systems*, 2020.

[17] M. Jourdan and A. Azize. Optimal best arm identification under differential privacy. *arXiv preprint arXiv:2510.17348*, 2025.

[18] M. Jourdan, R. Degenne, and E. Kaufmann. An epsilon-best-arm identification algorithm for fixed-confidence and beyond. *Advances in Neural Information Processing Systems*, 36: 16578–16649, 2023.

[19] E. Kaufmann, P. Ménard, O. Darwiche Domingues, A. Jonsson, E. Leurent, and M. Valko. Adaptive reward-free exploration. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132 of *Proceedings of Machine Learning Research*, pages 865–891, 2021.

[20] J. Lazzaro and C. Pike-Burke. Fixed-confidence multiple change point identification under bandit feedback. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

[21] J. Ok, A. Proutiere, and D. Tranos. Exploration in structured reinforcement learning. In *Advances in neural information processing systems*, volume 31. Curran Associates, Inc., 2018.

[22] R. Poiani, M. Bernasconi, and A. Celli. Non-asymptotic analysis of (sticky) track-and-stop. *arXiv preprint arXiv:2505.22475*, 2025.

[23] R. Poiani, M. Bernasconi, and A. Celli. Pure exploration with infinite answers. *arXiv preprint arXiv:2505.22473*, 2025.

[24] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[25] A. Russo and A. Pacchiano. Adaptive exploration for multi-reward multi-policy evaluation. *arXiv preprint arXiv:2502.02516*, 2025.

[26] A. Russo and A. Proutiere. Model-free active exploration in reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

[27] A. Russo and F. Vannella. Multi-reward best policy identification. In *Advances in Neural Information Processing Systems*, volume 37, pages 105583–105662, 2024.

[28] A. Russo, Y. Song, and A. Pacchiano. Pure exploration with feedback graphs. *arXiv preprint arXiv:2503.07824*, 2025.

[29] P. J. Schweitzer. Perturbation theory and finite markov chains. *Journal of Applied Probability*, 5(2), 1968.

[30] A. Sethi, K. S. Sagar, S. Agrawal, D. Basu, and P. Karthik. Asymptotically optimal sequential testing with markovian data. *arXiv preprint arXiv:2602.17587*, 2026.

[31] R. K. Sundaram. *A First Course in Optimization Theory*. Cambridge University Press, Cambridge, 1996.

[32] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[33] J. Taupin, Y. Jedra, and A. Proutiere. Best policy identification in linear mdps. In *2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1–8. IEEE, 2023.

[34] A. Tuynman and R. Degenne. The batch complexity of bandit pure exploration. In *ICML 2025-42nd International Conference on Machine Learning*, volume 267, pages 60442–60468, 2025.

[35] C. Zalinescu. Sharp estimates for hoffman's constant for systems of linear inequalities and equalities. *SIAM Journal on Optimization*, 2006.

Appendix

Contents

1 Introduction **1**

2 Problem Setting **2**

2.1 Sample Complexity Lower Bound 3

2.2 Assumptions 4

3 Navigate-and-Stop Algorithm **5**

3.1 Algorithm 6

4 Non-Asymptotic Upper Bound for NaS **7**

4.1 Sharpness of the Objective Function 7

4.2 Implicit and Explicit Upper Bounds 8

5 Related Work & Discussion **10**

A Further Related Work **14**

B Table of Notation **15**

C Proof for non-asymptotic upper bound **17**

C.1 Overview of proof 17

C.2 Good events 17

C.3 Consequences of good events 19

C.4 Tracking the limiting allocation 21

C.5 Information growth and the expectation bound 22

C.6 Correctness of the stopping rule 24

C.7 Asymptotic optimality and the restricted-class lower bound 25

C.8 Explicit bounds under polynomial sharpness 25

C.9 Proofs of Theorem 6 and Corollary 10 27

C.10 Supplementary results 29

C.10.1 Sufficient conditions for an irreducible target 29

C.10.2 Auxiliary optimization and policy results 30

D Sharpness of the Objective Function **33**

D.1 Main sharpness lemma 33

D.2 Verification of the selector conditions 34

D.3 Linear Sharpness Under a Finite Discretisation of the MDP Space 36

D.4 An explicit family realizing arbitrary rational exponents for polynomial sharpness 37

D.5 Empirical Verification of Sharpness Results. 43

E Other results **44**

A Further Related Work

Best Arm Identification & Sequential Testing in Bandits Methods for BPI are built upon ideas constructed in for optimal Best Arm Identification (BAI) and sequential testing in bandits. In particular, much of the work discussed here is a consequence of the seminal paper by Garivier and Kaufmann [13] which introduces the first asymptotically optimal method, Track-and-Stop. The main idea being to track the proportion of actions played across the action space with estimated optimal allocations (and then stop when sufficient evidence has been gathered). This blossomed into research studying problems with multiple correct answers [e.g. 11, 23], different objectives [e.g. 20, 18, 34, 17, 14], and different types of feedback [e.g. 28]. There has also been a line of work providing non-asymptotic analysis to these optimal methods. Notably Degenne et al. [12] consider game-theoretic variants of Track-and-Stop and provided complementary non-asymptotic bounds using the optimistic structure of the decisions. More recently, Poiani et al. [22] provide non-asymptotic analysis to the original Track-and-Stop algorithm, without making any optimistic modifications to the algorithm. Their theoretical techniques are what originally motivated the work in this paper on BPI. We note that the potential non-uniqueness of the true optimal allocations are well-handled in the bandit setting (both in the existing non-asymptotic and asymptotic analysis) due to 1-state nature of bandit problems and through the use of C-tracking. Whereas in the BPI setting of this paper, we have to be much more careful, this is discussed further in Section 3. In this work, we study the more broad RL setting of BPI.

Best Policy Identification In the full reinforcement learning setting we have both a state-space and action-space on which the agent needs to learn the dynamics of the MDP. Hence, although the existing algorithms for optimal BPI are strongly motivated by the existing work in bandits, the analysis can be significantly more involved and nuanced. We first note that there have been some optimal methods proposed in the setting under which the agent has access to a generative model which can provide observations from *any* state-action in every round [e.g. 1, 33]. An alternative approach, which we consider in this work, is to consider settings in which the agent only has access to a single trajectory through which they must navigate through the MDP. In Al Marjani et al. [2], they provide the first instance-dependent asymptotic guarantee for BPI with a single trajectory, through Navigate-and-Stop, motivated by Track-and-Stop [13]. Subsequently there have been optimal methods proposed for various other sequential testing problems, such as multi-reward settings [27, 25]. However, our work is the first to provide non-asymptotic guarantees for an optimal BPI method. In particular, we provide non-asymptotic guarantees for the Navigate-and-Stop variant analyzed here.

Other related works. Our work focuses on active sequential testing for BPI in online RL. A related but distinct line of work studies *policy testing*, where the goal is to decide whether a fixed policy has value above a prescribed threshold, while minimizing the sample complexity of the test [e.g. 3]. Concurrent work has also established non-asymptotic lower bounds for a broad class of sequential testing problems with Markovian data [30]. Their results, however, are derived for passive data-collection strategies, whereas BPI in our setting requires active exploration under navigation constraints; we discuss possible connections in later sections. Finally, regret-minimization objectives in RL have been extensively studied [e.g. 5, 21, 7], but they address a fundamentally different criterion from fixed-confidence identification, where the objective is to stop as quickly as possible while returning an optimal policy.

B Table of Notation

Notation	Definition
$\mathcal{Z}, S, A$	State-action space $\mathcal{S} \times \mathcal{A}$, with $S = \lvert\mathcal{S}\rvert$ and $A = \lvert\mathcal{A}\rvert$.
$\mathcal{M}, G_\mu(L)$	Known admissible class and common-support bounded transition class; all alternatives belong to $\mathcal{M} \subseteq G_\mu(L)$.
$\pi_o(M)$	Unique reward-maximizing deterministic policy of the true model.
$\mathcal{A}(\pi)$	Admissible models for which recommendation π is not optimal.
$\mathrm{Alt}(M)$	$\mathcal{A}(\pi_o(M))$ at an admissible model with a unique optimal policy.
$N_t(s,a), N_t(s)$	Counts among the t completed transitions before action t; $\sum_{s,a} N_t(s,a) = t$.
$\widehat{P}_t, P_t^{\Pi}$	Raw empirical transition kernel and its rowwise ℓ_1 projection onto $G_\mu(L)$.
$M_t, \widehat{\pi}_t^r$	Model with projected kernel, and its recommended policy $\widehat{\pi}_t^r = \pi_o(M_t)$ with fixed tie-breaking.
$J_N(w), J_N^*$	Information objective on the whole simplex, and its maximum over the navigation polytope $\Omega(N)$.
$T(M)$	Characteristic time $1/J_M^*$ relative to the fixed admissible class.
$\Omega^*(M), x$	Optimal allocation set and its minimum-norm member $x = \omega^*(M)$.
$\Omega_\eta(N)$	Allocations feasible under N with objective gap at most η.
$\omega_t^\star$	Minimum-norm member of $\Omega_{\eta_t}(M_t)$; it need not be exactly optimal for M_t.
r_t, η_t	Computable model confidence radius and oracle tolerance $\eta_t = \log(e+t)r_t$.
$\pi_t^{\mathrm{base}}, \pi_t$	Policy induced by $\omega_t^\star$, and the actual policy after mixing with forcing.
$\varepsilon_t(s), \pi_{f,t}$	State-dependent forcing weight and count-balancing forcing policy.
$\pi^\dagger, K^\dagger$	Policy induced by x, and its state transition kernel under M.
ρ_M	Minimum target state marginal $\min_s \sum_a x(s,a) > 0$.
D, m	Maximum shortest-path length in the state support graph and $m = D + 1$.
q, ζ	$q = 1 - \alpha$, $\zeta = q^m$: local forcing and global coverage exponents.
$p_*, \kappa, c_{\mathrm{cov}}$	Minimum positive true transition probability, $\kappa = p_*/(8A^2)$, and coverage coefficient in (9).
$\ell_t, h_t, b_{\mathrm{cov}}(t)$	Deterministic logarithmic factors and coverage envelope in (13).
E_t, ψ_t	Model/coverage good event with failure at most $2(t+1)^{-7}$, and its deterministic model-radius envelope.
$\Delta_M, r_{\mathrm{pol}}, r_{\mathrm{loc}}$	Minimum optimal-action gap, policy-stability radius, and radius for both labels and raw KL regularity in (15).
$L_M, L_\omega, L_{\mathrm{raw}}$	Objective regularity constants in Lemma 14.
φ_M	Pointwise sharpness modulus: distance to the true optimal set is at most φ_M of the information gap.
$\kappa_0, \kappa_1, \kappa_\Sigma$	Two one-sided allocation-transport constants from Lemma 31, and their sum.
C_M	Selector tolerance coefficient $2L_M + L_\omega \kappa_\Sigma$.
W_t, V_t, v_t	Allocation-error envelope, base-policy-error envelope, and actual base-policy error in (22).
h_z, H_M	Fixed Poisson solution for coordinate z, and maximum span of these solutions.
t_0, u_t	Initial period satisfying (20), and $u_t = \max\{t_0, \lceil\sqrt{t}\rceil\}$.
$\mathcal{R}_M(t)$	Count-error envelope in (27).
$\mathcal{E}_M(t)$	Information-error envelope in (28).

Notation	Definition (continued)
$e_t, \mathcal{D}_M(t)$	Power-logarithmic estimation envelope and simplified main-text penalty in (7); $\mathcal{E}_M(t) \le \mathcal{D}_M(t)$ for $t \ge t_M$ when $t_0 = t_M$.
$K_M, d_{S,A}, t_M$	Tracking coefficient, fluctuation coefficient, and initial time used in Theorem 6.
$K_\varphi, C_{\mathrm{stat}}$ $C_{\mathrm{force}}, C_{\mathrm{init}}$	Linear-sharpness allocation coefficient, combined statistical correction, forcing correction, and explicit initial-period cost in Corollary 10.
X_t, Z_t, τ_δ	Observed pair $X_t = (s_t, a_t)$, raw likelihood statistic in (5), and stopping time at confidence δ.
$b(N_t, \delta)$	Stopping threshold.
$c_1, c_2, \bar{b}$	Deterministic threshold constants and envelope: $c_1 = SA(S-1)$, $c_2 = c_1(1+\log 2)$, $\bar{b} = \log(1/\delta) + c_1 \log t + c_2$.
q_t, C_{tail}	Information-bound failure probability and its summable tail constant in (29)-(30).
p, C_φ	Polynomial-modulus parameters $\varphi_M(u) \le C_\varphi u^p$. Linear sharpness is $p = 1$.
a, k, K_W, K_V	Selector-rate exponent, logarithmic exponent, and coefficients in (36).
$\mathfrak{R}(C, d, k)$	Explicit cutoff ensuring $Ct^{-d}[\log(et)]^k \le 1$, defined in (32).
$A_0, A_{\mathrm{mart}}, A_{\mathrm{sel}}$ $A_{\mathrm{force}}, A_{\mathrm{mod}}$	Explicit information-error coefficients in (39).
B_δ, g, R, U_δ	Leading confidence scale and auxiliary quantities for the sharper appendix inversion in Corollary 21.

C Proof for non-asymptotic upper bound

C.1 Overview of proof

The proof follows the information collected along a single trajectory, from learning the dynamics to crossing the stopping threshold. Appendix C.2 first shows that forced exploration samples every transition row often enough for accurate estimation. Appendix C.3 then converts this model accuracy into proximity to a fixed information-optimal allocation and its induced policy. The geometric step is the stability of the relaxed minimum-norm selector, proved in Appendix D.1; Appendix D.2 verifies its assumptions using objective regularity and feasible-allocation perturbation bounds.

This stable selector is useful because near-optimal information values alone need not identify a single allocation when the optimizer is nonunique. Once a target is selected, its irreducibility allows the fixed Poisson argument in Appendix C.4 to convert policy errors into visitation errors. The proof tracks this target directly, without introducing an averaged oracle allocation or policy as an additional target. Time averages of the policy errors still enter the bound. This route relies on the relaxed selector and the target assumption of our specified NaS variant.

Appendix C.5 combines visitation and estimation errors into a lower bound on the stopping statistic, then sums failure probabilities to bound the expected stopping time. Correctness of the stopping rule is established separately in Appendix C.6. Appendix C.7 matches the leading coefficient to the lower bound, while Appendices C.8 and C.9 derive explicit polynomial-sharpness bounds and the simpler forms stated in the main paper. Appendix C.10 collects the supporting model-class and target-irreducibility results.

Two distinctions are useful throughout. The reward-maximizing policy $\pi_o(M)$ is different from the exploration policy $\pi^\dagger$ induced by the information-optimal allocation. Also, the projected kernel P_t^{Π} is used in the sampling oracle, whereas the stopping statistic uses the raw empirical kernel $\widehat{P}_t$. We establish correctness of that statistic independently of the tracking argument.

Filtration and indexing. Let $\mathcal{F}_t = \sigma(s_0, a_0, s_1, \ldots, a_{t-1}, s_t)$, enlarged by the algorithm's past randomization if necessary. The vector $\pi_t(\cdot|s_t)$ is $\mathcal{F}_t$-measurable. Conditional on $\mathcal{F}_t$ the action is sampled, and conditional on $\mathcal{F}_t \vee \sigma(a_t)$ the next state has distribution $P_M(\cdot|s_t, a_t)$. Counts refer to the t completed transitions before action t. For analysis, continue the sampling rule indefinitely after the stopping time; this extension agrees with the actual algorithm before stopping.

C.2 Good events

The sampling rule in Section 3 must provide accurate estimates even when the current allocation concentrates on only a few actions. We first use a predictable-mean Bernoulli bound to quantify how forced exploration spreads observations through the communicating support graph. Combining this coverage guarantee with concentration of the transition observations gives a confidence radius for both the raw and projected kernels. These two guarantees form the good events used in Appendix C.3 to control the estimated policy and allocation.

A time-uniform Bernoulli bound. The following time-uniform version of a Bernoulli bound, of the type used in Dann et al. [10], applies separately to the action-selection and transition-observation filtrations. Its predictable means allow the sampling decisions to depend on all observations available so far.

Lemma 11. *Let $X_j \in \{0,1\}$ be adapted observations with predictable means p_j. For every $h > 0$,*

$$\mathbb{P}\left(\exists n \geq 1 : \sum_{j<n} X_j < \tfrac{1}{2}\sum_{j<n} p_j - h\right) \leq e^{-h}.$$

Proof. Since $1 - p + p/e \leq \exp(-(1-e^{-1})p)$ and $1 - e^{-1} > 1/2$, we have $e^{p/2}(1 - p + p/e) \leq 1$. Thus $\exp(\sum_{j<n}(p_j/2 - X_j))$ is a nonnegative supermartingale starting at one. Ville's inequality gives the result, simultaneously over all n. □

Coverage along a single trajectory. The support graph has an edge $s \to u$ if $P_M(u|s,a) > 0$ for some action. Let D be its maximum shortest-path length, with distance zero from a state to itself, and set

$$m = D + 1, \qquad q = 1 - \alpha, \qquad \zeta = q^m,$$
$$p_* = \min\{P_M(u|s,a) : P_M(u|s,a) > 0\}, \qquad \kappa = \frac{p_*}{8A^2}, \tag{9}$$
$$c_{\mathrm{cov}} = \kappa^{(1-q^m)/(1-q)} S^{-\zeta}.$$

Here m counts at most D state transitions followed by one action selection along a shortest support path.

Lemma 12 (State-dependent coverage)**.** *For any adaptive base policies, $0 < \alpha < 1$, $\beta \geq 0$ and $\nu \in (0,1)$, with probability at least $1 - \lambda$, simultaneously for every integer $t \geq 1$ and every (s,a),*

$$N_t(s,a) \geq \left[c_{\mathrm{cov}} t^{\zeta} - 4 \left(\log \frac{S + S^2 A}{\lambda} + 1 \right) \right]_+ . \tag{10}$$

Proof. At each round j, choose $a_j^{\min}(s) \in \arg\min_a N_j(s,a)$ using a fixed tie-breaking rule before sampling a_j. This representative may change as counts increase. Let $R_t(s) = \sum_{j<t} \mathbb{1}\{s_j = s,\ a_j = a_j^{\min}(s)\}$ count its actual selections at state s. Recall from Algorithm 1 that $\varepsilon_j(s) = \max\{1+\nu, N_j(s)\}^{-\alpha}$. Conditional on $\mathcal{F}_j$ and $s_j = s$, the probability of selecting the representative is at least $\varepsilon_j(s)/A$, since a minimum-count action has forcing probability at least $1/A$. Lemma 11, with $h = \log((S + S^2A)/\lambda)$, gives

$$R_t(s) \geq \frac{1}{2A} \sum_{k=0}^{N_t(s)-1} \max\{1+\nu, k\}^{-\alpha} - h \qquad \text{for all } t$$

outside an event of probability e^{-h}. Deterministically, $R_t(s) \leq A(\min_a N_t(s,a) + 1)$: each counted play increments an action whose pre-count is at most the final minimum, and each action can undergo at most that minimum plus one such increments. For $n \geq 1$,

$$\sum_{k=0}^{n-1} \max\{1+\nu, k\}^{-\alpha} \geq \frac{n}{(n+1)^{\alpha}} \geq \tfrac{1}{2} n^q.$$

Including $n = 0$ and using $A \geq 1$ therefore yields

$$\min_a N_t(s,a) \geq \frac{N_t(s)^q}{4A^2} - (h+1). \tag{11}$$

For each positive transition (s,a,u), let $H_t(s,a,u)$ count its occurrences among the completed transitions. Applying the same Bernoulli bound after observing the action gives

$$H_t(s,a,u) \geq \tfrac{1}{2} P_M(u|s,a) N_t(s,a) - h \qquad \text{for all } t.$$

There can be one incoming visit at time t not counted in $N_t(u)$, so $N_t(u) \geq H_t(s,a,u) - 1$. Combining with (11) gives

$$N_t(u) \geq \kappa N_t(s)^q - 2(h+1).$$

A union bound over S local events and at most S^2A transition events makes these inequalities simultaneous with failure probability at most λ.

Put $B = 2(h+1)$ and $F(v) = [\kappa v^q - B]_+$. Every support edge, and the final choice of any action, give a count lower bound F. The function is increasing and $F(v) \leq v$: for $v \leq 1$ it is zero, and for $v \geq 1$ use $\kappa v^q \leq v$. At each time some state has at least t/S visits. Following a shortest support path from this state to the desired state, followed by the desired action, gives

$$N_t(s,a) \geq F^{\circ(d+1)}(t/S) \geq F^{\circ m}(t/S), \qquad d \leq D.$$

Here $F^{\circ j}$ denotes j successive compositions of F, with $F^{\circ 0}$ the identity; for example, $F^{\circ 2}(v) = F(F(v))$. The state and path may depend on the data because the edge inequalities already hold simultaneously. For $u, v \geq 0$ and $0 < q < 1$, $[u - v]_+^q \geq u^q - v^q$. Induction now gives

$$F^{\circ j}(v) \geq \kappa^{1+q+\cdots+q^{j-1}} v^{q^j} - \xi_j, \qquad \xi_0 = 0, \quad \xi_{j+1} = B + \kappa \xi_j^q.$$

Since $B \geq 2$ and $\kappa \leq 1/8$, the induction $\xi_j \leq 2B$ follows from $(2B)^q \leq 2B$. Substituting $j = m$ and using nonnegativity of the counts proves (10). □

The exponent $\zeta = (1-\alpha)^m$ is positive for every allowed α. It reflects the repeated propagation of coverage along a support path. The temperature β does not enter this bound because every minimum-count action has forcing probability at least $1/A$.

Concentration of empirical transition rows. We use a direct ℓ_1 concentration inequality and account for the factor of two introduced by projection. This also specifies the confidence radius used by the algorithm.

Lemma 13. *For every integer $t \geq 1$, the radius r_t in Algorithm 1 satisfies*

$$\mathbb{P}\left(\max\{\|\widehat{P}_t - P_M\|_{\infty,1}, \|P_t^{\Pi} - P_M\|_{\infty,1}\} > r_t\right) \leq (t+1)^{-7}.$$

Here we use the notation $\|Q\|_{\infty,1} := \max_{s,a} \sum_{s'} |Q(s'|s,a)|$ for the largest ℓ_1 norm across state-action rows.

Proof. The observations at successive visits to a fixed row can be coupled to an infinite i.i.d. sequence from that row. For its empirical law $\widehat{p}_n$,

$$\|\widehat{p}_n - p\|_1 = \max_{v \in \{-1,1\}^S} \langle v, \widehat{p}_n - p \rangle.$$

Each sign-projected observation has range length two. Hoeffding's inequality [6] and a union bound over the signs give

$$\mathbb{P}(\|\widehat{p}_n - p\|_1 > u) \leq 2^S e^{-nu^2/2}.$$

Union bound over all SA rows and the deterministic sample sizes $1 \leq n \leq t$ with $u = \sqrt{2\log(2^S SA\, t(t+1)^7)/n}$. Since the bound holds simultaneously for every $1 \leq n \leq t$, it also holds at the observed count $n = N_t(s,a)$ for each visited row. For each row, feasibility of the true row in the projection gives

$$\|P_t^{\Pi}(\cdot|s,a) - P_M(\cdot|s,a)\|_1 \leq 2\|\widehat{P}_t(\cdot|s,a) - P_M(\cdot|s,a)\|_1. \tag{12}$$

Both row distances are at most two regardless of data, which justifies clipping the resulting radius at two and the convention for unvisited rows. □

C.3 Consequences of good events

We now turn the coverage and concentration bounds of Lemmas 12 and 13 into control of the policies used by the algorithm. The deterministic, instance-dependent envelope ψ_t below bounds the computable radius r_t on a good event. Small model error first preserves the optimal-policy label and controls information values; selector stability then yields allocation and base-policy error bounds. The selector and feasible-set arguments are proved in Appendices D.1 and D.2, and the resulting policy bound is the input to tracking in Appendix C.4.

Define

$$\ell_t = \log(2^S SA\, t(t+1)^7), \qquad h_t = \log((S+S^2A)(t+1)^7),$$
$$b_{\mathrm{cov}}(t) = [c_{\mathrm{cov}} t^{\zeta} - 4(h_t+1)]_+, \qquad \psi_t = \begin{cases} 2, & b_{\mathrm{cov}}(t) = 0, \\ \min\{2, \sqrt{8\ell_t / b_{\mathrm{cov}}(t)}\}, & b_{\mathrm{cov}}(t) > 0. \end{cases} \tag{13}$$

Let E_t be the intersection of the concentration event in Lemma 13 and the coverage event in Lemma 12 with $\lambda = (t+1)^{-7}$. Then

$$\mathbb{P}(E_t^c) \leq 2(t+1)^{-7}, \qquad \max\{\|\widehat{P}_t - P_M\|_{\infty,1}, \|P_t^{\Pi} - P_M\|_{\infty,1}\} \leq r_t \leq \psi_t \quad \text{on } E_t. \tag{14}$$

In particular $\psi_t = O_M(\sqrt{\log(et)}\, t^{-\zeta/2})$, where O_M hides constants depending on the fixed instance and algorithm parameters, but not on t or δ. The failure probabilities are summable, so the Borel-Cantelli lemma makes these bounds valid eventually almost surely. Recall that $\eta_t = \log(e+t) r_t$ is the algorithm's tolerance for near-optimal information allocations. The logarithmic multiplier eventually dominates the fixed sensitivity constant needed by the selector argument, while $\log(e+t)\psi_t \to 0$ ensures that the tolerance still vanishes. Thus both model estimates converge and $\eta_t \to 0$ almost surely.

Policy labels and objective regularity. Let

$$\Delta_M = \min_{s,\, a \neq \pi_o(M)(s)} [V_M^*(s) - Q_M^*(s,a)] > 0,$$
$$R_r = \max\{1, \|r\|_\infty\}, \qquad r_{\mathrm{pol}} = \frac{\Delta_M (1-\gamma)^2}{4\gamma R_r}, \qquad r_{\mathrm{loc}} = \min\{r_{\mathrm{pol}}, p_*/2\}. \tag{15}$$

The following lemma fixes the alternative set before applying a KL perturbation estimate.

Lemma 14. *If $\|P_N - P_M\|_{\infty,1} \le r_{\mathrm{pol}}$, the unique optimal policy of N is $\pi_o(M)$. If additionally $\mathrm{Alt}(M) \neq \varnothing$ and $P_N \in G_\mu(L)$, then on the whole allocation simplex,*

$$|J_N(\omega) - J_M(\omega)| \le L_M \|P_N - P_M\|_{\infty,1}, \qquad L_M = 1 + 2\log L, \tag{16}$$
$$|J_M(\omega) - J_M(v)| \le L_\omega \|\omega - v\|_\infty, \qquad L_\omega = 2SA \log L. \tag{17}$$

Moreover, if a raw empirical kernel $\widehat{P}$ is supported on the common support and $\|\widehat{P} - P_M\|_{\infty,1} \le p_*/2$, then

$$\sup_{Q \in G_\mu(L)} \max_{s,a} |\mathrm{KL}(\widehat{P}(\cdot|s,a), Q(\cdot|s,a)) - \mathrm{KL}(P_M(\cdot|s,a), Q(\cdot|s,a))| \le L_{\mathrm{raw}} \|\widehat{P} - P_M\|_{\infty,1}, \tag{18}$$

where $L_{\mathrm{raw}} = 1 + \log(2L^2)$.

Proof. Bellman contraction gives

$$\|V_N^* - V_M^*\|_\infty \le \frac{\gamma R_r}{(1-\gamma)^2} \|P_N - P_M\|_{\infty,1}, \qquad \|Q_N^* - Q_M^*\|_\infty \le \frac{\gamma R_r}{(1-\gamma)^2} \|P_N - P_M\|_{\infty,1}.$$

Within r_{pol}, each optimal Q-value changes by at most $\Delta_M/4$. Comparing the original optimal action with any competitor therefore reduces their gap by at most $2(\Delta_M/4)$, leaving a gap of at least $\Delta_M/2 > 0$. Thus $\pi_o(N) = \pi_o(M)$ remains the unique optimal policy. Since $\mathcal{A}(\pi)$ depends only on the policy π and the fixed admissible class, this also gives $\mathcal{A}(\pi_o(N)) = \mathrm{Alt}(M)$, even if $N \notin \mathcal{M}$. The information objectives can consequently be compared over the same alternatives.

On the common support, $p_i, q_i \in [\mu_i/L, L\mu_i]$, so $|1 + \log(p_i/q_i)| \le 1 + 2\log L$ and $0 \le \mathrm{KL}(p,q) \le 2\log L$. The mean-value theorem along the row segment proves (16) after taking the infimum over the same alternatives. The bound on KL coefficients proves (17), since $\|\omega - v\|_1 \le SA\|\omega - v\|_\infty$. For the final assertion, every positive coordinate of P_M is at least p_*. Along the segment to $\widehat{P}$, each coordinate lies between half and twice its value under P_M. Consequently its ratio to the corresponding coordinate of any $Q \in G_\mu(L)$ is between $1/(2L^2)$ and $2L^2$. The same derivative argument gives (18). Coordinates outside the common support contribute zero throughout. □

Selector and policy error. Let $x = \omega^*(M)$ and $\rho_M = \min_s x(s) > 0$. Changing the transition kernel changes both the information objective and the feasible stationary allocations. Lemma 31 in Appendix D.2 controls the latter effect with fixed instance-dependent constants κ_0, κ_1: every allocation feasible under N is within $\kappa_0 \|P_N - P_M\|_{\infty,1}$ of $\Omega(M)$, and the particular target x is within $\kappa_1 \|P_N - P_M\|_{\infty,1}$ of $\Omega(N)$. The first constant is a Hoffman constant for the true navigation polytope; the second measures the stationary-distribution sensitivity of the fixed target policy.

Lemma 30 combines these bounds with Lemma 14 and pointwise sharpness to control the selected allocation. The near-optimal tolerance admits a transported copy of x, and minimum-norm selection then gives proximity to this particular target even when the optimal set is not a singleton. To express the resulting error envelope, set

$$\kappa_\Sigma = \kappa_0 + \kappa_1, \qquad C_M = 2L_M + L_\omega \kappa_\Sigma, \qquad W_t = \sqrt{\kappa_\Sigma \psi_t + \varphi_M((\log(e+t) + C_M)\psi_t)}. \tag{19}$$

Choose an integer $t_0 \ge 1$ such that, for every $t \ge t_0$,

$$\psi_t \le r_{\mathrm{loc}}, \qquad \log(e+t) \ge C_M, \qquad W_t \le \rho_M/(2A). \tag{20}$$

The three conditions ensure, respectively, stable policy labels and raw-KL regularity, a tolerance large enough relative to model perturbations, and stable normalization of allocations into action probabilities. Such a finite t_0 exists because $\psi_t \log(e+t) \to 0$ and $\varphi_M(u) \to 0$ as $u \downarrow 0$. An explicit

choice under polynomial sharpness is given in Appendix C.8, equation (37). Define the base-policy error envelope

$$V_t = \min\left\{2, \frac{A(A+1)}{\rho_M} W_t\right\} \quad (t \geq t_0). \tag{21}$$

On E_t, $\eta_t = \log(e+t) r_t \geq C_M \|P_t^{\Pi} - P_M\|_{\infty,1}$. Lemma 30 therefore yields

$$\|\omega_t^\star - x\|_\infty \leq W_t, \qquad v_t := \max_s \|\pi_t^{\text{base}}(\cdot|s) - \pi^\dagger(\cdot|s)\|_1 \leq V_t. \tag{22}$$

Thus W_t controls allocation error and V_t controls the corresponding base-policy error. Both vanish with t, and the latter bound will control cumulative visitation error in the next subsection.

C.4 Tracking the limiting allocation

Appendix C.3 controls how closely the base policy follows the fixed target policy $\pi^\dagger$. We now show how this translates into visitation frequencies close to x. We use the Poisson-equation martingale method also employed in the NaS analysis [2], with a single correction for the fixed target chain. The resulting decomposition separates a boundary term, random fluctuations, and accumulated deviations of the adaptive policy from $\pi^\dagger$. A separate pathwise bound controls forced exploration, providing the count estimate needed for information growth in Appendix C.5.

For $z \in \mathcal{S} \times \mathcal{A}$, let $f_z(s,a) = \mathbb{1}\{(s,a) = z\}$ and $f_z^\dagger(s) = \sum_a \pi^\dagger(a|s) f_z(s,a)$. Let $d(s) = x(s)$. By Assumption 3, $K^\dagger$ is irreducible with unique stationary distribution d, and $\sum_s d(s) f_z^\dagger(s) = x(z)$. To center accumulated visits at this stationary frequency, introduce a state-dependent correction h_z solving the Poisson equation

$$h_z - K^\dagger h_z = f_z^\dagger - x(z)\mathbf{1}, \qquad H_M = \max_z \operatorname{span}(h_z) < \infty. \tag{23}$$

Here $\operatorname{span}(h_z) = \max_s h_z(s) - \min_s h_z(s)$, so H_M is the largest range of these corrections over all state-action pairs. For a finite irreducible chain, every centered right-hand side has a Poisson solution. In this case one is $h_z = (I - K^\dagger + \mathbf{1}d)^{-1}(f_z^\dagger - x(z)\mathbf{1})$. To verify the inverse, if $(I - K^\dagger + \mathbf{1}d)v = 0$, multiplication by d gives $dv = 0$, and then $K^\dagger v = v$. Irreducibility makes such a harmonic vector constant, hence zero. The centering identity then verifies the displayed solution. Any two solutions differ by a constant, so their spans agree. This argument does not require aperiodicity.

The finite constant H_M depends on the target chain and controls its transient visitation error. One interpretation is through hitting times under the fixed target policy. For an auxiliary chain (Y_k) with kernel $K^\dagger$, let $T_{s'} = \inf\{k \geq 0 : Y_k = s'\}$. Summing the Poisson equation up to this hitting time gives

$$h_z(s) - h_z(s') = \mathbb{E}_s^{K^\dagger}\left[\sum_{k=0}^{T_{s'}-1} \left(f_z^\dagger(Y_k) - x(z)\right)\right], \qquad H_M \leq \max_{s,s'} \mathbb{E}_s^{K^\dagger}[T_{s'}].$$

Indeed, first stop at $T_{s'} \wedge n$ and take expectations in the telescoping Poisson identity. Finite irreducibility gives finite expected hitting times, and boundedness of h_z and of the summand justifies the limit as $n \to \infty$. The inequality follows since the summand has absolute value at most one. These hitting times use the fixed target policy, whereas the MDP diameter optimizes over policies. We retain the potentially smaller quantity H_M in the bounds.

Proposition 15 (Fixed-target tracking)**.** *For every fixed integer $t \geq 1$ and $\lambda \in (0,1)$, with probability at least $1 - \lambda$,*

$$\left\| \frac{N_t}{t} - x \right\|_\infty \leq \frac{H_M}{t} + \frac{1 + H_M}{2t} \sum_{j<t} d_j + (1 + H_M)\sqrt{\frac{\log(2SA/\lambda)}{2t}}. \tag{24}$$

Here $d_j = \|\pi_j(\cdot|s_j) - \pi^\dagger(\cdot|s_j)\|_1$. This probability statement is unconditional.

Proof. Write $P_a h(s) = \sum_u P_M(u|s,a)h(u)$ and define

$$B_{j,z} = \sum_a (\pi_j - \pi^\dagger)(a|s_j)[f_z(s_j,a) + P_a h_z(s_j)],$$
$$D_{j+1,z} = f_z(s_j,a_j) + h_z(s_{j+1}) - \mathbb{E}[f_z(s_j,a_j) + h_z(s_{j+1}) \mid \mathcal{F}_j].$$

Since h_z is a fixed function of the true instance, $D_{j+1,z}$ is $\mathcal{F}_{j+1}$-measurable and satisfies $\mathbb{E}[D_{j+1,z} \mid \mathcal{F}_j] = 0$, even though the policy is adaptive. Equation (23) and the definition of $B_{j,z}$ give

$$\mathbb{E}[f_z(s_j, a_j) + h_z(s_{j+1}) \mid \mathcal{F}_j] = x(z) + h_z(s_j) + B_{j,z}.$$

Subtract this conditional mean and sum over $j < t$. The correction terms telescope, yielding the exact decomposition

$$N_t(z) - tx(z) = h_z(s_0) - h_z(s_t) + \sum_{j<t} D_{j+1,z} + \sum_{j<t} B_{j,z}.$$

The boundary term satisfies $|h_z(s_0) - h_z(s_t)| \le \operatorname{span}(h_z) \le H_M$. The bracket in $B_{j,z}$ has range at most $1 + H_M$, and the action-probability differences sum to zero, so $|B_{j,z}| \le (1 + H_M)d_j/2$. Conditional on $\mathcal{F}_j$, the variable defining $D_{j+1,z}$ has range length at most $1 + H_M$. The conditional Hoeffding bound [6] thus gives

$$\mathbb{P}\left(\left|\sum_{j<t} D_{j+1,z}\right| > (1 + H_M)\sqrt{\tfrac{t}{2}\log(2SA/\lambda)}\right) \le \lambda/(SA).$$

A union bound over z, followed by division by t, proves the proposition. □

Proposition 16 (Cumulative forcing)**.** *Every trajectory satisfies*

$$\sum_{j<t} \varepsilon_j(s_j) \le S + \frac{S^\alpha}{1-\alpha} t^{1-\alpha}. \tag{25}$$

Proof. On the $(k+1)$st visit to a state its local count is k. For $n \ge 1$,

$$\sum_{k=0}^{n-1} \max\{1 + \nu, k\}^{-\alpha} \le 1 + \sum_{k=1}^{n-1} k^{-\alpha} \le 1 + \frac{n^{1-\alpha}}{1-\alpha}.$$

Sum over states and use concavity: $\sum_s N_t(s)^{1-\alpha} \le S^\alpha t^{1-\alpha}$. □

The sampling mixture gives $d_j \le v_j + 2\varepsilon_j(s_j)$. Combining this inequality with Propositions 15 and 16 yields, with probability at least $1 - \lambda$,

$$\|N_t - tx\|_\infty \le H_M + (1+H_M)\left(S + \frac{S^\alpha}{1-\alpha} t^{1-\alpha} + \frac{1}{2}\sum_{j<t} v_j\right) + (1+H_M)\sqrt{\tfrac{t}{2}\log(2SA/\lambda)}. \tag{26}$$

The forced-exploration contribution has cumulative power $1 - \alpha$. Connectivity enters the estimation and selector errors through the coverage exponent ζ, rather than through this cumulative forcing cost.

C.5 Information growth and the expectation bound

We now combine the estimation guarantees of Appendix C.3 with the visitation bound of Appendix C.4. Lemma 14 converts these two errors into a loss of information relative to the ideal rate $1/T(M)$. The resulting lower bound on Z_t holds with summable failure probabilities, which allow us to control the entire tail of the stopping time and hence its expectation. This gives the implicit bound used by the asymptotic and explicit analyses in Appendices C.7 to C.9.

Recall the deterministic base-policy error envelope V_j from (21), with $v_j \le V_j$ on E_j by (22). For an integer $t \ge t_0$, put $u_t = \max\{t_0, \lceil\sqrt{t}\rceil\}$, and define the deterministic envelopes

$$\mathcal{R}_M(t) = H_M + (1 + H_M)\left[u_t + S + \frac{S^\alpha}{1-\alpha} t^{1-\alpha} + \frac{1}{2}\sum_{j=u_t}^{t-1} V_j\right] + (1 + H_M)\sqrt{\tfrac{t}{2}\log(2SA(t+1)^3)}, \tag{27}$$

$$\mathcal{E}_M(t) = L_\omega \mathcal{R}_M(t) + L_{\mathrm{raw}}\, t\psi_t. \tag{28}$$

Empty sums are zero. For averages over all earlier rounds, set $V_j = 2$ for $0 \le j < t_0$, the largest possible ℓ_1 distance between action distributions. The count envelope $\mathcal{R}_M(t)$ includes the early rounds, forced exploration, accumulated policy errors, and martingale fluctuations. In (28), $L_\omega \mathcal{R}_M(t)$ is the information loss from imperfect visitation, while $L_{\mathrm{raw}} t\psi_t$ is the loss from transition-estimation error. The constants $L_\omega, L_{\mathrm{raw}}$ are given by Lemma 14, equations (17) and (18), and ψ_t is defined in (13). The factor t converts the latter per-observation error into accumulated information loss.

Proposition 17. *For every integer $t \ge t_0$,*

$$\mathbb{P}\left(Z_t < \frac{t}{T(M)} - \mathcal{E}_M(t) \right) \le q_t, \qquad q_t = \frac{1}{3t^3} + (t+1)^{-3}. \tag{29}$$

Moreover $\mathcal{E}_M(t) = o(t)$.

Proof. Intersect E_j for $u_t \le j \le t$. From (14), the failure probability is at most

$$2 \sum_{j=u_t}^{\infty} (j+1)^{-7} \le 2 \int_{u_t}^{\infty} x^{-7}\, dx = \frac{1}{3u_t^6} \le \frac{1}{3t^3}.$$

On this event, $v_j \le V_j$ for $u_t \le j < t$, whereas $v_j \le 2$ always. Hence $\sum_{j<t} v_j \le 2u_t + \sum_{j=u_t}^{t-1} V_j$. Apply (26) with $\lambda = (t+1)^{-3}$ and intersect its event afterward. This gives $\|N_t - tx\|_\infty \le \mathcal{R}_M(t)$ outside total failure probability q_t.

On E_t, for $t \ge t_0$, Lemma 14 and (20) ensure that M_t has the same unique optimal policy as M. Consequently, $\widehat{\pi}_t^r = \pi_o(M)$, and the alternative set $\mathcal{A}(\widehat{\pi}_t^r)$ used in Z_t equals $\mathrm{Alt}(M)$. With $w = N_t/t$, apply the raw-KL and allocation-continuity bounds of Lemma 14, equations (18) and (17), to obtain

$$\begin{aligned} Z_t/t &\ge J_M(w) - L_{\mathrm{raw}}\psi_t \\ &\ge J_M(x) - L_\omega \|w - x\|_\infty - L_{\mathrm{raw}}\psi_t = T(M)^{-1} - L_\omega \|w - x\|_\infty - L_{\mathrm{raw}}\psi_t. \end{aligned}$$

The bounds apply to w because Lemma 14 holds on the whole allocation simplex. This proves (29). Finally $u_t/t \to 0$, $V_j \to 0$, and $\psi_t \to 0$. The arithmetic averages $t^{-1}\sum_{j<t} V_j$ tend to zero: the finitely many initial terms contribute $O(1/t)$, and the remaining terms are eventually arbitrarily small. Together with $\alpha > 0$ and the vanishing martingale term after division by t, this proves $\mathcal{E}_M(t)/t \to 0$. □

Theorem 18 (Implicit finite-confidence bound)**.** *Let $c_1 = SA(S-1)$, $c_2 = c_1(1 + \log 2)$, and*

$$\bar{b}(t,\delta) = \log(1/\delta) + c_1 \log t + c_2,$$

$$n_\delta = \inf\left\{ n \in \mathbb{N} : n \ge t_0,\ \forall t \in \mathbb{N},\ t \ge n :\ \frac{t}{T(M)} - \mathcal{E}_M(t) > \bar{b}(t,\delta) \right\}.$$

Thus n_δ is the earliest deterministic time, at least t_0, after which the information lower envelope exceeds the threshold upper bound at every later integer time. This ensures that the stopping-time tail can be bounded at every index being summed. Under Assumptions 1-3, $n_\delta < \infty$ and

$$\mathbb{E}_M[\tau_\delta] \le n_\delta + C_{\mathrm{tail}}, \qquad C_{\mathrm{tail}} := \sum_{t=1}^{\infty} q_t = \frac{4}{3}\sum_{t=1}^{\infty} t^{-3} - 1 < 1. \tag{30}$$

Proof. For $t \ge 1$, $N_t(s,a) \le t$ and $S \ge 2$ give $b(N_t,\delta) \le \bar{b}(t,\delta)$. Proposition 17 gives $\mathcal{E}_M(t) = o(t)$; together with logarithmic growth of $\bar{b}(t,\delta)$, this implies $n_\delta < \infty$. For every $t \ge n_\delta$, Proposition 17, equation (29), and the stopping rule give

$$\mathbb{P}_M(\tau_\delta > t) \le \mathbb{P}_M(Z_t < t/T(M) - \mathcal{E}_M(t)) \le q_t.$$

Consequently

$$\mathbb{E}[\tau_\delta] = \sum_{t=0}^{\infty} \mathbb{P}(\tau_\delta > t) \le n_\delta + \sum_{t=n_\delta}^{\infty} q_t \le n_\delta + C_{\mathrm{tail}}.$$

For the last inequality in (30), $\sum_{t\ge 1} t^{-3} < 1 + \int_1^\infty x^{-3} dx = 3/2$. □

C.6 Correctness of the stopping rule

The preceding analysis bounds the time needed to collect information; we now verify that crossing the stopping threshold returns the correct policy with probability at least $1-\delta$. We use the transition-likelihood mixture argument of Al Marjani et al. [2, Lemma 15], written explicitly for adaptive row visits; related confidence constructions appear in Jonsson et al. [16], Kaufmann et al. [19]. We then combine coverage from Lemma 12 with policy stability from Lemma 14 to establish finite expected stopping even without target irreducibility. This proves correctness on the full admissible class used by the lower bound in Appendix C.7.

Lemma 19 (Error control and finite stopping)**.** *Under Assumptions 1-2, the algorithm is δ-PC on $\mathcal{M}$. In fact, for each $M \in \mathcal{M}$, it stops almost surely with finite expectation, and*

$$\mathbb{P}_M(\tau_\delta < \infty,\ \widehat{\pi}^r_{\tau_\delta} \notin \Pi^*(M)) \le \delta.$$

This assertion does not require Assumption 3.

Proof. **Error control.** For each row $z = (s, a)$, let k_z be its known support size. Mix the trajectory likelihood ratios against P_M over independent uniform Dirichlet priors on these row supports. For a fixed candidate kernel Q on this support, its likelihood ratio is

$$L_t(Q) = \prod_{j<t} \frac{Q(s_{j+1}|s_j, a_j)}{P_M(s_{j+1}|s_j, a_j)}.$$

Its conditional expectation at the next step equals its present value, since the action probabilities cancel and the next-state ratio has conditional mean one. Integrating gives a nonnegative martingale $\mathcal{L}_t$ with $\mathcal{L}_0 = 1$. The mixing kernels need not lie in $G_\mu(L)$; they are only used to construct this test martingale.

If a row has counts n_i with $n = \sum_i n_i$, its mixed likelihood is $(k_z-1)!\prod_i n_i!/(n+k_z-1)!$. Its maximum empirical likelihood is $\prod_i (n_i/n)^{n_i}$, with the value one when $n = 0$. Since the multinomial probability at its empirical parameter is at most one, their ratio is at most $\binom{n+k_z-1}{k_z-1}$. Multiplication over rows therefore gives

$$\exp\left(\sum_z N_t(z)\mathrm{KL}(\widehat{P}_t(\cdot|z), P_M(\cdot|z))\right) \le \mathcal{L}_t \prod_z \binom{N_t(z)+k_z-1}{k_z-1}.$$

For $k_z = 1$ the binomial coefficient is one. For all rows it is bounded by $\binom{N_t(z)+S-1}{S-1} \le [e(1 + N_t(z)/(S-1))]^{S-1}$. Ville's inequality now proves

$$\mathbb{P}_M\left(\exists t \ge 1 : \sum_z N_t(z)\mathrm{KL}(\widehat{P}_t(\cdot|z), P_M(\cdot|z)) \ge b(N_t, \delta)\right) \le \delta.$$

If the policy returned at time $t = \tau_\delta$ is not optimal in M, then $M \in \mathcal{A}(\widehat{\pi}^r_t)$ and the sum in this display is at least $Z_t \ge b(N_t, \delta)$. This proves the error bound for the implemented statistic.

Finite stopping on the full admissible class. We establish finite expected stopping for every $M \in \mathcal{M}$, including instances where the selected target is reducible. If $\mathrm{Alt}(M) = \varnothing$, then once the estimated optimal policy satisfies $\widehat{\pi}^r_t = \pi_o(M)$, its alternative set is empty and $Z_t = +\infty$. The estimation events in (14) ensure this eventually, with summable failure probabilities.

Otherwise, local policy stability implies that every $Q \in \mathrm{Alt}(M)$ is at rowwise distance greater than r_{pol} from P_M. On E_t, for all sufficiently large deterministic t, the selected policy satisfies $\widehat{\pi}^r_t = \pi_o(M)$ and $\|\widehat{P}_t - P_M\|_{\infty,1} \le r_{\mathrm{pol}}/2$. Thus at least one row has $\|\widehat{P}_t(\cdot|z) - Q(\cdot|z)\|_1 > r_{\mathrm{pol}}/2$. Pinsker's inequality [9], in natural logarithms, gives

$$\sum_z N_t(z)\mathrm{KL}(\widehat{P}_t(\cdot|z), Q(\cdot|z)) \ge \frac{r^2_{\mathrm{pol}}}{8} \min_z N_t(z).$$

Taking the infimum preserves this bound. By (10), on E_t the right-hand side eventually exceeds $\overline{b}(t, \delta)$, because t^ζ grows faster than $\log t$ and $\zeta > 0$. Hence there is a finite deterministic threshold after which $\mathbb{P}_M(\tau_\delta > t) \le 2(t+1)^{-7}$. This proves both almost-sure stopping and finite expectation without using target irreducibility. The same tail argument applies to the empty-alternative case. □

C.7 Asymptotic optimality and the restricted-class lower bound

The information-growth bound in Proposition 17 has a sublinear error, suggesting the leading sample-complexity coefficient $T(M)$. We first specialize the navigation-constrained change-of-measure lower bound to the same known-reward class on which Lemma 19 establishes correctness. We then apply Theorem 18 to obtain the matching upper coefficient. This proves Theorem 8 while keeping the model class fixed in both directions. The usual change-of-measure argument [2, 13] gives, for $0 < \delta < 1/2$, any algorithm that is δ-PC on $\mathcal{M}$ and every $N \in \mathrm{Alt}(M)$,

$$\sum_z \mathbb{E}_M[N_{\tau_\delta}(z)]\mathrm{KL}(P_M(\cdot|z), P_N(\cdot|z)) \geq \mathrm{kl}(1-\delta, \delta).$$

For finite expected stopping time this follows by the likelihood-ratio chain rule at the stopping time and data processing for the event that the returned policy equals $\pi_o(M)$; otherwise the lower bound is immediate. Bounded log ratios in the common class justify the stopped expectation. Since the alternative set is nonempty and all row KL coefficients are at most $2\log L$, $\mathbb{E}_M[\tau_\delta] \to \infty$ as $\delta \downarrow 0$.

Let $w_\delta(z) = \mathbb{E}_M[N_{\tau_\delta}(z)]/\mathbb{E}_M[\tau_\delta]$. The expected flow imbalance at each state is at most one, by telescoping the state indicators. Hence every subsequential limit of w_δ lies in $\Omega(M)$. By the uniform Lipschitz bound (17), J_M is continuous on the entire simplex. Along a subsequence realizing the lower limit of expected sample complexity divided by $\log(1/\delta)$, take a further convergent subsequence of w_δ. The change-of-measure inequalities imply

$$J_M(w_\delta) \geq \frac{\mathrm{kl}(1-\delta, \delta)}{\mathbb{E}_M[\tau_\delta]}.$$

Since the limiting feasible objective is at most $J_M^* = 1/T(M)$ and $\mathrm{kl}(1-\delta, \delta)/\log(1/\delta) \to 1$, this proves Theorem 2 for the specified class.

For the matching upper bound, fix $\epsilon > 0$ and put $a_\delta = (1+\epsilon)T(M)\log(1/\delta)$. Uniformly over integers $t \geq \lceil a_\delta \rceil$, $\log(1/\delta)/t \leq 1/((1+\epsilon)T(M))$, whereas $\sup_{t \geq a_\delta}(\mathcal{E}_M(t) + c_1 \log t + c_2)/t \to 0$. The all-later-times condition in Theorem 18 therefore holds from $\lceil a_\delta \rceil$ onward for all sufficiently small δ. It follows that

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}_M[\tau_\delta]}{\log(1/\delta)} \leq (1+\epsilon)T(M).$$

Letting $\epsilon \downarrow 0$, and using Lemma 19 and the matching lower bound, proves Theorem 8.

C.8 Explicit bounds under polynomial sharpness

Theorem 18 bounds the expected stopping time through an information envelope, but its stopping horizon remains implicit. Here we assume polynomial sharpness and use the estimation and selector bounds from Appendix C.3 to give explicit powers of time for each error contribution. A cutoff construction and the global inversion lemma below then turn these estimates into a finite-confidence bound, retaining the initial cost as well as the leading confidence term. Appendix C.9 uses the same ingredients to derive the simpler main-text forms. Specifically, assume

$$\varphi_M(u) \leq C_\varphi u^p \quad \text{for all } u \geq 0, \qquad C_\varphi > 0,\ p > 0. \tag{31}$$

A local bound of this form can be extended globally by increasing C_φ, because distances within the simplex are bounded. Here p is the exponent of φ_M itself; linear sharpness means $p = 1$, and the selector bound contains its square root.

An explicit cutoff operator. For $C, d > 0$ and $k \geq 0$, define

$$\mathfrak{R}(C, d, k) = u\,[4(1+k/d)^2 \log(eu)]^{k/d}, \qquad u = \max\{1, C^{1/d}\}. \tag{32}$$

When $k = 0$ the bracket has exponent zero. This operator has the useful property

$$t \geq \mathfrak{R}(C, d, k) \quad \Longrightarrow \quad C\,t^{-d}[\log(et)]^k \leq 1. \tag{33}$$

To verify it, put $q = k/d$, $l = \log(eu) \geq 1$, $c = 4(1+q)^2$, and $R = u(cl)^q$. The elementary bound $1 + q\log c + q \leq c$ implies $\log(eR) = l + q\log c + q\log l \leq cl$. For completeness, $\log c = \log 4 + 2\log(1+q) \leq 2 + 2q$, so $1 + q\log c + q \leq 1 + 3q + 2q^2 \leq 4(1+q)^2$. Thus $R^d \geq C[\log(eR)]^k$. Also $\log(eR) \geq q$ because $\log c \geq \log 4 > 1$. The logarithmic derivative of $t^{-d}[\log(et)]^k$ is $[-d + k/\log(et)]/t$, so it is nonincreasing beyond R. This proves (33) for every real $t \geq R$.

Explicit initial period and error constants. Let

$$\begin{aligned} c_\ell &= S\log 2 + \log(SA) + 8, & c_h &= 4[\log(S + S^2A) + 8], \\ C_\psi &= \sqrt{16c_\ell/c_{\mathrm{cov}}}, & t_{\mathrm{cov}} &= \mathfrak{R}(2c_h/c_{\mathrm{cov}}, \zeta, 1). \end{aligned} \tag{34}$$

For $t \geq 1$, $\ell_t \leq c_\ell \log(et)$ and $4(h_t + 1) \leq c_h \log(et)$. Therefore, for $t \geq t_{\mathrm{cov}}$, $b_{\mathrm{cov}}(t) \geq c_{\mathrm{cov}} t^\zeta/2$ and

$$\psi_t \leq C_\psi t^{-\zeta/2}[\log(et)]^{1/2}. \tag{35}$$

Set

$$\begin{aligned} a &= \frac{\zeta}{4}\min\{1, p\}, & k &= \frac{1}{4}\max\{1, 3p\}, \\ K_W &= \sqrt{\kappa_\Sigma C_\psi} + \sqrt{C_\varphi}\,[(2 + C_M)C_\psi]^{p/2}, & K_V &= \frac{A(A+1)}{\rho_M} K_W. \end{aligned} \tag{36}$$

Indeed $\log(e + t) \leq 2\log(et)$ for $t \geq 1$. Apply $\sqrt{x + y} \leq \sqrt{x} + \sqrt{y}$ in (19) and use (31) to obtain

$$W_t \leq \sqrt{\kappa_\Sigma C_\psi}\, t^{-\zeta/4}[\log(et)]^{1/4} + \sqrt{C_\varphi}[(2+C_M)C_\psi]^{p/2} t^{-\zeta p/4}[\log(et)]^{3p/4} \leq K_W t^{-a}[\log(et)]^k.$$

In particular, the feasible-set transport contribution must be retained when $p > 1$. Choose

$$\begin{aligned} t_0 = \lceil \max\{1, t_{\mathrm{cov}},\ &\mathfrak{R}(C_\psi/r_{\mathrm{loc}}, \zeta/2, 1/2), \\ &e^{C_M}, \mathfrak{R}(2AK_W/\rho_M, a, k)\}\rceil. \end{aligned} \tag{37}$$

Equations (33) and (35) verify every condition in (20) for all $t \geq t_0$. On the good events, the policy error is bounded by $K_V t^{-a}[\log(et)]^k$.

For $0 < a < 1$,

$$\sum_{j=u_t}^{t-1} V_j \leq K_V[\log(et)]^k \sum_{j=1}^{t-1} j^{-a} \leq \frac{K_V}{1-a} t^{1-a}[\log(et)]^k.$$

Using $u_t \leq t_0 + 1 + \sqrt{t}$ and $\log(2SA(t+1)^3) \leq [\log(2SA) + 3]\log(et)$ in (28), we obtain the nondecreasing envelope

$$\begin{aligned} \mathcal{E}_M(t) \leq \overline{\mathcal{E}}_M(t) := A_0 &+ A_{\mathrm{mart}} t^{1/2}[\log(et)]^{1/2} \\ &+ A_{\mathrm{sel}} t^{1-a}[\log(et)]^k + A_{\mathrm{force}} t^{1-\alpha} \\ &+ A_{\mathrm{mod}} t^{1-\zeta/2}[\log(et)]^{1/2}, \qquad t \geq t_0. \end{aligned} \tag{38}$$

with the explicit constants

$$\begin{aligned} A_0 &= L_\omega[H_M + (1 + H_M)(t_0 + 1 + S)], \\ A_{\mathrm{mart}} &= L_\omega(1 + H_M)\left[1 + \sqrt{(\log(2SA) + 3)/2}\right], \\ A_{\mathrm{sel}} &= \frac{L_\omega(1 + H_M)K_V}{2(1-a)}, \qquad A_{\mathrm{force}} = \frac{L_\omega(1 + H_M)S^\alpha}{1-\alpha}, \\ A_{\mathrm{mod}} &= L_{\mathrm{raw}} C_\psi. \end{aligned} \tag{39}$$

All powers in this display are strictly below one. The selector power is $1 - a$ with $a = \zeta\min\{1, p\}/4$, and the forcing power is $1 - \alpha$. Thus the dominant power is $r = 1 - \min\{\alpha, a\}$. For linear sharpness it is $1 - \min\{\alpha, (1-\alpha)^m/4\}$. Logarithmic factors and the initial period are specified in (38) and (37).

Why a global inversion is needed. An inequality $t \leq B + Ct^r$, $0 \leq r < 1$, may have solutions on the scale $C^{1/(1-r)}$ even when B is small. Evaluating the correction only at B omits this scale. For instance, when $r = 1/2$, the largest solution equals $[(C + \sqrt{C^2 + 4B})/2]^2$. The following argument retains the initial scale while keeping leading coefficient one for B.

Lemma 20 (Global inversion)**.** *Let $B \geq 1$ and let $g : [1, \infty) \to [0, \infty)$ be nondecreasing. If $g(t) \leq t/2$ for every $t \geq R$, then*

$$t \leq B + g(t) \quad \Longrightarrow \quad t \leq \max\{R, B + g(2B)\}.$$

Proof. For $t < R$ there is nothing to prove. Otherwise $t \le B + t/2$, so $t \le 2B$ and monotonicity gives $t \le B + g(2B)$. □

Corollary 21 (Polynomial sharpness)**.** *Under Assumptions 1-3 and* (31)*, take* t_0 *from* (37)*. Let*

$$B_\delta = \max\{1, T(M)[\log(1/\delta) + c_2]\}, \qquad g(t) = T(M)c_1 \log(et) + T(M)\overline{\mathcal{E}}_M(t).$$

Write $g(t) = \sum_{i=1}^{6} C_i t^{r_i} [\log(et)]^{k_i}$*. The table separates the six sources of error and specifies the coefficients and exponents used in the cutoff below:*

Contribution	C_i	r_i	k_i
Threshold	$T(M)c_1$	0	1
Initial period and boundary	$T(M)A_0$	0	0
Martingale and initial rounds	$T(M)A_{\mathrm{mart}}$	$1/2$	$1/2$
Selector	$T(M)A_{\mathrm{sel}}$	$1-a$	k
Forcing	$T(M)A_{\mathrm{force}}$	$1-\alpha$	0
Current model estimation	$T(M)A_{\mathrm{mod}}$	$1-\zeta/2$	$1/2$

Define

$$R = \max_{1 \le i \le 6} \mathfrak{R}(12C_i, 1 - r_i, k_i), \qquad U_\delta = \max\{t_0, R, B_\delta + g(2B_\delta)\}. \tag{40}$$

Then the explicit finite-confidence bound is

$$\mathbb{E}_M[\tau_\delta] \le \lfloor U_\delta \rfloor + 1 + C_{\mathrm{tail}}. \tag{41}$$

For fixed instance quantities and algorithm parameters, the correction above $T(M)\log(1/\delta)$ *is* $\widetilde{O}_M((\log(1/\delta))^{1-\min\{\alpha,a\}})$ *as* $\delta \downarrow 0$*, where the notation hides these fixed constants and powers of* $\log\log(1/\delta)$*. The finite-confidence bound* (41) *specifies them explicitly.*

Proof. By (33), every summand of $g(t)$ is at most $t/12$ for all $t \ge R$, so $g(t) \le t/2$. Suppose an integer $t \ge t_0$ fails the sufficient stopping inequality based on $\overline{\mathcal{E}}_M$ and $\overline{b}$. Then

$$t \le T(M)[\log(1/\delta) + c_2] + T(M)c_1 \log t + T(M)\overline{\mathcal{E}}_M(t) \le B_\delta + g(t).$$

Lemma 20 implies $t \le \max\{R, B_\delta + g(2B_\delta)\}$. Therefore every integer $t \ge \lfloor U_\delta \rfloor + 1$ satisfies the strict stopping inequality. The expectation bound follows from (30). The coefficients C_i, exponents r_i, k_i, and initial cutoffs t_0, R are independent of δ. Since every $r_i < 1$, this gives the stated fixed-instance expansion. That expansion describes the small-δ regime; the uniform finite-confidence assertion is (41), including t_0 and R. □

C.9 Proofs of Theorem 6 and Corollary 10

We derive the main-text forms from the information envelope of Appendix C.5 and the explicit estimates and global inversion lemma in Appendix C.8. First, we choose the constants and initial time so that the previous estimation and selector guarantees apply, then prove $\mathcal{E}_M(t) \le \mathcal{D}_M(t)$ and invoke Theorem 18. Under linear sharpness, we group the smaller contributions with the allocation error and apply Lemma 20 to obtain Corollary 10. These slightly looser bounds display the main sources of finite-confidence cost while retaining the same leading coefficient.

Estimation constants and initial time. Let p_* be the smallest positive transition probability and put

$$c_{\mathrm{cov}} = \Big(\frac{p_*}{8A^2}\Big)^{[1-(1-\alpha)^m]/\alpha} S^{-\zeta},\ C_\psi = 4\sqrt{\frac{S\log 2 + \log(SA) + 8}{c_{\mathrm{cov}}}},\ c_h = 4[\log(S + S^2A) + 8].$$

Choose an integer $t_M \ge 1$ such that, for every real $t \ge t_M$,

$$\begin{gathered} c_{\mathrm{cov}} t^\zeta \ge 2c_h \log(et), \qquad e_t \le r_{\mathrm{loc}}, \qquad \log(e+t) \ge C_M, \\ \kappa_\Sigma e_t + \varphi_M\big((\log(e+t) + C_M)e_t\big) \le \frac{\rho_M^2}{4A^2}. \end{gathered} \tag{42}$$

As in (15), r_{loc} is the local radius for policy stability and raw-KL regularity. These conditions ensure estimation and allocation stability apply together. They hold eventually since $e_t \log(e+t) \to 0$ and $\varphi_M(u) \to 0$ as $u \downarrow 0$; an explicit linear-sharpness choice is given below.

Explicit initial cost under linear sharpness. Recall the cutoff function $\mathfrak{R}$ from (32) and its property (33). With $a = \zeta/4$ and K_φ as in Corollary 10, an admissible initial time and the corresponding initial-period cost are

$$\begin{aligned} t_M = \lceil \max\{1, \mathfrak{R}(2c_h/c_{\mathrm{cov}}, \zeta, 1), \mathfrak{R}(C_\psi/r_{\mathrm{loc}}, \zeta/2, 1/2), \\ e^{C_M}, \mathfrak{R}(2AK_\varphi/\rho_M, a, 3/4)\} \rceil , \\ C_{\mathrm{init}} = 2 + T(M)K_M(t_M + S + 2) \\ + \max\{t_M, 8T(M)K_M(t_M + S + 2), \mathfrak{R}(8T(M)c_1, 1, 1), \\ \mathfrak{R}(8T(M)C_{\mathrm{stat}}, a, 3/4), \mathfrak{R}(8T(M)C_{\mathrm{force}}, \alpha, 0)\} . \end{aligned} \tag{43}$$

Theorem 6. Take $t_0 = t_M$ in the sharper appendix bounds. The first condition in (42), together with (34), ensures $b_{\mathrm{cov}}(t) \geq c_{\mathrm{cov}} t^\zeta/2$ and hence $\psi_t \leq e_t$ for every $t \geq t_M$. Monotonicity of φ_M and the other conditions in (42) verify (20). For $j \geq t_M$, therefore,

$$V_j \leq \frac{A(A+1)}{\rho_M} \sqrt{\kappa_\Sigma e_j + \varphi_M\big((\log(e+j) + C_M)e_j\big)}.$$

Extend the nonnegative sum from $j = u_t$ to $j = 1$. Since $u_t \leq t_M + 1 + \sqrt{t}$, $L_\omega H_M \leq K_M$, and

$$\sqrt{\frac{t}{2}\log(2SA(t+1)^3)} \leq \sqrt{\frac{\log(2SA)+3}{2}}\sqrt{t\log(et)},$$

substitution in (27)-(28) proves

$$\mathcal{E}_M(t) \leq \mathcal{D}_M(t), \qquad t \geq t_M.$$

No estimation claim is needed for the added early summands: they merely enlarge the bound. Both e_t and $\varphi_M((\log(e+t) + C_M)e_t)$ tend to zero. Their arithmetic averages therefore tend to zero, and $0 < \alpha < 1$ gives $\mathcal{D}_M(t) = o(t)$. Thus its all-later-times stopping horizon is finite and at least the one in Theorem 18. Replacing $C_{\mathrm{tail}} < 1$ by 1 proves (6).

The linear-sharpness penalty. For every real $j \geq 1$, $\log(e+j) \leq 2\log(ej)$ and $\log(ej) \geq 1$. If $\varphi_M(u) \leq C_\varphi u$, then

$$\sqrt{\kappa_\Sigma e_j + \varphi_M((\log(e+j) + C_M)e_j)} \leq K_\varphi j^{-a}[\log(ej)]^{3/4}, \qquad a = \zeta/4. \tag{44}$$

The cutoff property (33) applied to each term in the displayed definition of t_M verifies all four conditions in (42). In particular, $t \geq e^{C_M}$ ensures the logarithmic condition, and the final cutoff makes the right-hand side of (44) at most $\rho_M/(2A)$.

For integers $t \geq 1$, use $0 < a < 1$ to obtain

$$\sum_{j=1}^{t-1} j^{-a}[\log(ej)]^{3/4} \leq \frac{t^{1-a}[\log(et)]^{3/4}}{1-a}.$$

Also $1/2 \leq 1-a$, $1-\zeta/2 \leq 1-a$, and $[\log(et)]^{1/2} \leq [\log(et)]^{3/4}$. Consequently, for $t \geq t_M$,

$$\mathcal{D}_M(t) \leq K_M(t_M + S + 2) + C_{\mathrm{stat}} t^{1-a}[\log(et)]^{3/4} + C_{\mathrm{force}} t^{1-\alpha}. \tag{45}$$

This bound groups the current-estimation and martingale contributions with the slower allocation term. Their individual powers are given in (38).

Corollary 10. For this proof only, put $D_0 = K_M(t_M + S + 2)$ and define

$$G(t) = T(M)\left[c_1\log(et) + D_0 + C_{\mathrm{stat}} t^{1-a}[\log(et)]^{3/4} + C_{\mathrm{force}} t^{1-\alpha}\right], \qquad t \geq 1.$$

Let Q denote the maximum in the definition of C_{init}, so $C_{\mathrm{init}} = 2 + T(M)D_0 + Q$. Each of the four nonnegative terms in $G(t)$ is at most $t/8$ for every $t \geq Q$: this follows from the three cutoff-function terms and the condition $t \geq 8T(M)D_0$. Thus $G(t) \leq t/2$. Moreover, G is nondecreasing on $[1, \infty)$.

If an integer $t \geq t_M$ fails the sufficient stopping inequality with the right-hand side of (45), then $t \leq B_\delta + G(t)$. Lemma 20 yields $t \leq \max\{Q, B_\delta + G(2B_\delta)\}$. Every later integer therefore satisfies that stopping inequality, and the sharper expectation bound gives

$$\mathbb{E}_M[\tau_\delta] \leq \lfloor \max\{Q, B_\delta + G(2B_\delta)\} \rfloor + 1 + C_{\mathrm{tail}} \leq Q + B_\delta + G(2B_\delta) + 2.$$

Expanding $G(2B_\delta)$ and collecting its constant term with $Q + 2$ proves (8). Every constant in C_{init} is independent of δ, so the additive relaxation preserves leading coefficient $T(M)$.

C.10 Supplementary results

This subsection supplies the structural facts used in the preceding analysis. The product-polytope structure of $G_\mu(L)$ justifies rowwise projection, and bounded KL coefficients give objective regularity and a positive finite characteristic time. Appendix C.10.1 then gives sufficient conditions for the target irreducibility used in tracking and feasible-allocation transport. The remaining optimization and reward-policy facts are collected separately in Appendix C.10.2.

Lemma 22. *The set $G_\mu(L)$ is a nonempty compact convex product of row polytopes.*

Proof. The reference kernel μ belongs to $G_\mu(L)$. Each row is the intersection of the probability simplex with the closed coordinate bounds $\mu_i/L \le Q_i \le L\mu_i$, hence is a nonempty compact convex polytope. The row constraints do not couple distinct state-action pairs, proving the product assertion. In particular, any convex combination of two admissible kernels preserves normalization, the known zero coordinates, and all bounds. □

Lemma 23. *For probability rows p_1, p_2, q satisfying the same $G_\mu(L)$ row constraints,*

$$|\mathrm{KL}(p_1, q) - \mathrm{KL}(p_2, q)| \le (1 + 2\log L)\|p_1 - p_2\|_1, \qquad 0 \le \mathrm{KL}(p_1, q) \le 2\log L.$$

Proof. Restrict to the common positive support. Along the segment between p_1 and p_2, the derivative of $p_i \log(p_i/q_i)$ is $1 + \log(p_i/q_i)$, whose absolute value is bounded by $1 + 2\log L$. The mean-value theorem in the ℓ_1-ℓ_∞ dual norms gives the first assertion. The bound $p_i/q_i \le L^2$ gives the second after summing with weights p_i. Zero-support coordinates contribute zero and are omitted from differentiation. □

Lemma 24 (Nondegenerate characteristic time)**.** *If $M \in \mathcal{M}$ is communicating, has a unique optimal policy, and $\mathrm{Alt}(M) \neq \varnothing$, then $0 < J_M^* \le 2\log L < \infty$. The objective is continuous and concave on the whole simplex, and its maximizing set in $\Omega(M)$ is nonempty, compact, and convex.*

Proof. The uniform-action policy induces an irreducible state chain by communication. Its stationary state-action allocation w_u has every coordinate positive. Policy stability from Lemma 14 implies that every alternative is at rowwise ℓ_1 distance greater than r_{pol} from M. Pinsker's inequality on at least one row therefore yields

$$J_M(w_u) \ge \tfrac{1}{2} \min_z w_u(z)\, r_{\mathrm{pol}}^2 > 0.$$

Every alternative row divergence is at most $2\log L$, proving finiteness. Continuity follows from the uniform bound on the linear coefficients, and concavity follows by taking the infimum of linear functions. The navigation polytope is nonempty and compact, so the maximum is attained; its set of maximizers is closed and convex. □

C.10.1 Sufficient conditions for an irreducible target

Assumption 3 requires irreducibility only of the policy induced by the selected optimum. Full positivity of optimal allocations is sufficient, but is not necessary. In particular, the structured sharpness examples in Appendix D.4 have zero optimal action weights and fully supported transition rows, so their target chains are still irreducible.

Lemma 25 (Admissible one-row alternatives)**.** *Suppose that for every $z = (s, a)$ there exists $M^{(z)} \in \mathrm{Alt}(M)$ with $P_{M^{(z)}}(\cdot|u, b) = P_M(\cdot|u, b)$ at every $(u, b) \neq z$. Then every optimal allocation $w \in \Omega^*(M)$ has $w(z) > 0$ for all z. Consequently, communication implies that its induced state chain is irreducible.*

Proof. If $w(z) = 0$, testing the information objective against $M^{(z)}$ gives

$$J_M(w) \le \sum_{u,b} w(u, b)\mathrm{KL}(P_M(\cdot|u, b), P_{M^{(z)}}(\cdot|u, b)) = 0.$$

All row divergences are finite by the common bounded class. Lemma 24 gives $J_M^* > 0$, so such a w cannot be optimal. If all its coordinates are positive, its induced policy gives every action positive probability, and its state kernel contains every support edge of the communicating MDP. It is therefore irreducible. Under the aperiodicity assumption it is also aperiodic, although that is not needed in our tracking proof. □

The following sufficient criterion uses value gaps to construct policy-changing perturbations within the admissible model class. Let $V = V_M^{\pi_o(M)}$, $\Delta(s,a) = V(s) - Q_M^{\pi_o(M)}(s,a)$, and choose $s_+ \in \arg\max_s V(s)$ and $s_- \in \arg\min_s V(s)$. The maximum here is $\max_s V(s)$, not $\|V\|_\infty$ when rewards may be negative.

Proposition 26 (Value-gap criterion with admissible perturbations)**.** *For each suboptimal pair (s,a), suppose there is a $\lambda_{s,a} \in (0,1)$ such that the model obtained by replacing only its row with*

$$Q(\cdot|s,a) = (1-\lambda_{s,a})P_M(\cdot|s,a) + \lambda_{s,a}\delta_{s_+}$$

belongs to $\mathcal{M}$ and satisfies

$$\gamma\lambda_{s,a}\big(\max_u V(u) - P_M(\cdot|s,a)^\top V\big) > \Delta(s,a). \tag{46}$$

For each optimal row $(s, \pi_o(M)(s))$, suppose there is a $\lambda_s \in (0,1)$ such that replacing only that row with

$$Q(\cdot|s,\pi_o(M)(s)) = (1-\lambda_s)P_M(\cdot|s,\pi_o(M)(s)) + \lambda_s\delta_{s_-}$$

gives a model in $\mathcal{M}$ and

$$\gamma\lambda_s(1-\gamma)\big(P_M(\cdot|s,\pi_o(M)(s))^\top V - \min_u V(u)\big) > \min_{a\neq\pi_o(M)(s)} \Delta(s,a). \tag{47}$$

Then the hypothesis of Lemma 25 holds.

Proof. **Changing a suboptimal row.** Write $\pi = \pi_o(M)$. Since its own transition rows have not changed, the value of π under the perturbed model remains V. At the changed pair the advantage relative to $V(s)$ is exactly

$$-\Delta(s,a) + \gamma\lambda_{s,a}(V(s_+) - P_M(\cdot|s,a)^\top V) > 0.$$

Thus π is not optimal in this admissible model, which is an alternative differing in one row.

Changing an optimal row. Fix s_0 and let K' be the transition matrix of π under the perturbed model. Let $b = P_M(\cdot|s_0,\pi(s_0))^\top V - V(s_-) \geq 0$ and $W = (I-\gamma K')^{-1}e_{s_0}$. Subtracting the two Bellman equations gives

$$V_Q^\pi - V = -\gamma\lambda_{s_0} bW.$$

The vector W is the discounted expected number of visits to s_0. It is nonnegative, $W(s_0) \geq 1$, and $W(s) \leq W(s_0)$ for all s. Indeed, if its maximum were attained strictly away from s_0, the Bellman identity $W(s) = \gamma\sum_u K'(s,u)W(u)$ there would contradict positivity of the maximum. For any $a \neq \pi(s_0)$, its row was not changed, and its advantage under Q is

$$\begin{aligned} Q_Q^\pi(s_0,a) - V_Q^\pi(s_0) &= -\Delta(s_0,a) + \gamma\lambda_{s_0} b\,[W(s_0) - \gamma P_M(\cdot|s_0,a)^\top W] \\ &\geq -\Delta(s_0,a) + \gamma\lambda_{s_0} b(1-\gamma). \end{aligned}$$

Choose an action with minimum gap and use (47); its advantage is strictly positive. Hence this admissible model is also a one-row alternative. Applying the argument to every row proves the claim. □

The admissibility checks in this proposition cannot be dropped. In particular, moving mass to s_+ or s_- may violate a known zero, an upper bound, or a lower bound in $G_\mu(L)$. The numerical gap inequalities with $\lambda = 1$ only ensure enough improvement along an unrestricted ray; they do not ensure that an allowed point on that ray achieves it. The main theorem uses the weaker direct target condition rather than requiring these sufficient inequalities.

C.10.2 Auxiliary optimization and policy results

The following results concern the existence and geometry of optimal allocations and a separate property of reward-optimal policies. Berge's theorem provides general optimization background; Lemma 28 also gives a direct proof of the properties needed to define the minimum-norm selector.

Theorem 27 (Berge's Maximum Theorem [4, 31])**.** *Consider two sets $\mathcal{X} \subset \mathbb{R}^n$ and $\mathcal{Y} \subset \mathbb{R}^m$. Let $f : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ be a continuous function and let $\psi : \mathcal{X} \to \mathcal{P}(\mathcal{Y})$ be a non-empty compact-valued continuous correspondence. Define*

1. $f^\star(x) = \max\{f(x,y) \mid y \in \psi(x)\}$.
2. $\psi^\star(x) = \{y \in \psi(x) \mid f(x,y) = f^\star(x)\}$.

Then

1. *The function $f^\star$ is a continuous function on $\mathcal{X}$ and $\psi^\star$ is a compact-valued upper-hemi-continuous correspondence on $\mathcal{X}$.*
2. *If f is concave on $\mathcal{X} \times \mathcal{Y}$, and ψ has a convex graph, then $f^\star$ is a concave function and $\psi^\star$ is a convex-valued upper-hemi-continuous correspondence.*

Lemma 28. *For a fixed model with nonempty alternatives, the set of optimizers $\Omega^*(M)$ is nonempty, compact, and convex, and J_M is continuous and concave on the whole allocation simplex.*

Proof. This also follows directly without a continuity assumption on a moving alternative correspondence. Write $k_N(s,a) = \mathrm{KL}(P_M(\cdot|s,a), P_N(\cdot|s,a))$. These coefficients are in $[0, 2\log L]$. For $\lambda \in [0,1]$,

$$\begin{aligned} J_M((1-\lambda)w + \lambda v) &= \inf_{N \in \mathrm{Alt}(M)} [(1-\lambda)\langle w, k_N\rangle + \lambda\langle v, k_N\rangle] \\ &\geq (1-\lambda)J_M(w) + \lambda J_M(v). \end{aligned}$$

Thus J_M is concave. The bounded coefficients also give $|J_M(w) - J_M(v)| \leq 2SA\log L\|w - v\|_\infty$. It is therefore continuous, so it attains its maximum on the compact navigation polytope; the maximizer set is closed and convex. Equivalently, one may apply Berge's theorem to the closure of the alternatives inside $G_\mu(L)$: this closure is compact and taking it does not change the infimum. □

The following separate fact concerns reward-optimal policies. It does not imply that a uniform information-optimal exploration policy is impossible.

Lemma 29. *Let (S, A, P, r, γ) be a finite discounted Markov decision process (MDP) with $0 < \gamma < 1$ and $|A| \geq 2$. Suppose that the* uniform policy

$$\pi_u(a|s) = \frac{1}{|A|}, \quad \forall s \in S,\, a \in A$$

is optimal. Then π_u is not unique; in fact, every policy is optimal.

Proof. Let V^π denote the value function of policy π:

$$V^\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \,\middle|\, s_0 = s\right].$$

Define the optimal value and Q-functions by

$$V^*(s) = \max_{a \in A} Q^*(s,a), \qquad Q^*(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a)V^*(s').$$

Step 1. Characterization of optimal policies. A standard result in dynamic programming (see, e.g., Puterman [24]) states that a policy π is optimal if and only if, for all $s \in S$,

$$\pi(a|s) > 0 \;\Rightarrow\; Q^*(s,a) = V^*(s). \tag{48}$$

In words: a policy may assign positive probability only to actions that achieve the maximal $Q^*(s,a)$ value in each state.

Step 2. Apply (48) **to the uniform policy.** Assume that the uniform policy π_u is optimal. Since $\pi_u(a|s) = 1/|A| > 0$ for every s and a, substituting into (48) gives

$$Q^*(s,a) = V^*(s), \quad \forall s \in S,\, a \in A. \tag{49}$$

Thus, in each state, all actions have identical optimal Q-values.

Step 3. The set of optimal actions. Define the set of optimal actions in state s as

$$A^*(s) = \{a \in A : Q^*(s,a) = V^*(s)\}.$$

Equation (49) implies

$$A^*(s) = A, \quad \forall s \in S.$$

Hence, every action is optimal in every state.

Step 4. The set of optimal policies. Let Π^* denote the set of optimal policies:

$$\Pi^* = \big\{\pi : \pi(a|s) > 0 \Rightarrow a \in A^*(s),\ \forall s \in S\big\}.$$

Since $A^*(s) = A$ for all s, we have

$$\Pi^* = \{\pi : \pi(\cdot|s) \in \Delta(A),\ \forall s \in S\},$$

where $\Delta(A)$ is the simplex over A. That is, *every possible policy* (deterministic or stochastic) is optimal.

Step 5. Non-uniqueness. Because Π^* contains all policies, the uniform policy π_u is not unique. Indeed, any policy π' differing from π_u on at least one (s,a) is also optimal.

Step 6. Degenerate exception. If $|A| = 1$, the MDP admits only one policy. In that trivial case the uniform policy is uniquely optimal, but this is vacuous.

Conclusion. For any finite discounted MDP with $|A| > 1$, if the uniform policy is optimal, it cannot be the unique optimal policy. □

D Sharpness of the Objective Function

In this section we discuss the formal sharpness results used in the main paper and the sample-complexity analysis. First, we show how a shrinking near-optimal set gives a stable minimum-norm allocation selector using sharpness only at the true model. We then justify the transport bounds and objective assumptions in the MDP setting. Finally, finite-alternative and continuum examples illustrate how the shape of the allocation objective affects convergence.

D.1 Main sharpness lemma

For any nonempty set C, define $\mathrm{dist}_\infty(w, C) = \inf_{z\in C} \|w - z\|_\infty$. We work with the information objective J_N on the whole simplex, as defined in Section 3. Write $O(N) = \arg\max_{w\in\Omega(N)} J_N(w)$ and $x = \omega^*(M)$.

Lemma 30 (Stability of the relaxed minimum-norm selector)**.** *Suppose the alternatives at M are nonempty and the following conditions hold for the model N under consideration, with $e_N = \|P_N - P_M\|_{\infty,1}$. Assume the navigation sets are nonempty compact convex sets and the information objectives are finite, continuous, and concave. Require:*

(A2) On the whole simplex, $|J_N(w) - J_M(w)| \le L_M e_N$ and $|J_M(w) - J_M(v)| \le L_\omega \|w - v\|_\infty$.

(A3) The two one-sided transport bounds hold:

$$\sup_{y\in\Omega(N)} \mathrm{dist}_\infty(y, \Omega(M)) \le \kappa_0 e_N, \qquad \mathrm{dist}_\infty(x, \Omega(N)) \le \kappa_1 e_N.$$

(A4) At the true model only, $\mathrm{dist}_\infty(v, O(M)) \le \varphi_M(J_M^* - J_M(v))$ for $v \in \Omega(M)$, where φ_M is nondecreasing, $\varphi_M(0) = 0$, and $\lim_{u\downarrow 0} \varphi_M(u) = 0$.

Let $C_M = 2L_M + L_\omega(\kappa_0 + \kappa_1)$ and

$$y = \arg\min_{w\in\Omega_\eta(N)} \tfrac{1}{2}\|w\|_2^2, \qquad \Omega_\eta(N) = \{w \in \Omega(N) : J_N(w) \ge J_N^* - \eta\}.$$

If $\eta \ge C_M e_N$, then

$$\|y - x\|_\infty \le \sqrt{(\kappa_0 + \kappa_1)e_N + \varphi_M(\eta + C_M e_N)}. \tag{50}$$

If additionally $\rho_M = \min_s x(s) > 0$ and $\|y - x\|_\infty \le \rho_M/(2A)$, the induced policies satisfy

$$\max_s \|\pi_y(\cdot|s) - \pi^\dagger(\cdot|s)\|_1 \le \frac{A(A+1)}{\rho_M}\|y - x\|_\infty. \tag{51}$$

Proof. The objectives are concave, as infima of linear functions of the allocation, and continuous under (A2) and the bounded-class regularity. Their superlevel sets within the compact convex navigation polytopes are compact and convex. Hence the minimum-norm selectors exist and are unique.

Step 1: compare optimal values. Transport the particular optimum x to some $\bar{x} \in \Omega(N)$ with $\|\bar{x} - x\|_\infty \le \kappa_1 e_N$. Applying (A2) twice gives

$$J_N^* \ge J_N(\bar{x}) \ge J_M^* - (L_M + L_\omega\kappa_1)e_N.$$

Conversely, transport any maximizer of J_N to $\Omega(M)$ using the first bound in (A3). This yields

$$J_N^* \le J_M^* + (L_M + L_\omega\kappa_0)e_N.$$

Together these bounds imply $J_N^* - J_N(\bar{x}) \le C_M e_N \le \eta$, so $\bar{x} \in \Omega_\eta(N)$. This inclusion is the reason for using a near-optimal set rather than an exact tie-break.

Step 2: locate the selected point near the true optimal set. Choose $\bar{y} \in \Omega(M)$ with $\|y - \bar{y}\|_\infty \le \kappa_0 e_N$. Since $J_N(y) \ge J_N^* - \eta$,

$$J_M^* - J_M(\bar{y}) \le \eta + C_M e_N.$$

By (A4) and compactness of $O(M)$, there is a point $z \in O(M)$ satisfying

$$\|z - y\|_\infty \le \kappa_0 e_N + \varphi_M(\eta + C_M e_N).$$

Only the modulus at M is used.

Step 3: use the two projection inequalities. The minimum-norm points are Euclidean projections of zero onto their respective convex sets. Since $z \in O(M)$ and $\bar{x} \in \Omega_\eta(N)$,

$$\langle x, z - x\rangle \geq 0, \qquad \langle y, \bar{x} - y\rangle \geq 0.$$

As $\|x\|_1 = \|y\|_1 = 1$,

$$\begin{aligned}\|x - y\|_2^2 &\leq \langle x, z - y\rangle + \langle y, \bar{x} - x\rangle \\ &\leq \|z - y\|_\infty + \|\bar{x} - x\|_\infty \\ &\leq (\kappa_0 + \kappa_1)e_N + \varphi_M(\eta + C_M e_N).\end{aligned}$$

Taking square roots and using $\|x - y\|_\infty \leq \|x - y\|_2$ proves (50).

Step 4: normalize rows. Let $e = \|y - x\|_\infty$. Then $y(s) \geq x(s) - Ae \geq \rho_M/2 > 0$. The identity

$$\frac{y(s,a)}{y(s)} - \frac{x(s,a)}{x(s)} = \frac{y(s,a) - x(s,a)}{x(s)} + \frac{y(s,a)}{y(s)}\frac{x(s) - y(s)}{x(s)}$$

gives an entrywise bound $(A+1)e/\rho_M$. Summing over actions proves (51). □

Why pointwise sharpness alone does not stabilize exact tie-breaking. On $X = \{(u, 1-u) : 1/4 \leq u \leq 3/4\}$ consider $J_\theta(u, 1-u) = 1 + \theta u$. At $\theta = 0$ every allocation is optimal and the minimum-norm optimum is $(1/2, 1/2)$. For every $\theta > 0$ the unique optimum is $(3/4, 1/4)$. Thus exact minimum-norm selection need not be continuous, although each fixed objective has a sharpness modulus. A common modulus vanishing at zero over nearby θ would require $1/4 \leq \varphi(\theta/4)$ for every $\theta > 0$, which is impossible. This example explains the logical distinction; it is not asserted to be an MDP counterexample.

D.2 Verification of the selector conditions

Objective regularity (A2). Lemma 14 proves (A2) with $L_M = 1 + 2\log L$ and $L_\omega = 2SA\log L$ whenever $\mathrm{Alt}(M) \neq \varnothing$, $\|P_N - P_M\|_{\infty,1} \leq r_{\mathrm{pol}}$, and $P_N \in G_\mu(L)$. The restriction to a policy-stable neighborhood is necessary because the alternative set depends on the optimal-policy label.

Lemma 31 (One-sided feasible-allocation transport)**.** *Fix M and let $x \in \Omega(M)$ induce an irreducible state kernel $K^\dagger$ with stationary row vector $d(s) = x(s)$. For every N whose kernel lies in $G_\mu(L)$, the transport inequalities (A3) hold with a fixed Hoffman constant κ_0 of $\Omega(M)$ and*

$$\kappa_1 = \|(I - K^\dagger + \mathbf{1}d)^{-1}\|_{\infty\to\infty}.$$

Neither constant depends on N.

Proof. **Transport from $\Omega(N)$ to $\Omega(M)$.** Write the fixed navigation polytope as

$$\begin{aligned}\Omega(M) &= \{w \geq 0 : \mathbf{1}^\top w = 1,\ B(P_M)w = 0\}, \\ [B(P)w](s) &= \sum_a w(s,a) - \sum_{u,a} P(s|u,a)w(u,a).\end{aligned}$$

For $y \in \Omega(N)$, nonnegativity and normalization have zero residual, and

$$\begin{aligned}\|B(P_M)y\|_\infty &= \|(B(P_M) - B(P_N))y\|_\infty \\ &\leq \sum_{u,a} y(u,a) \max_s |P_M(s|u,a) - P_N(s|u,a)| \leq e_N.\end{aligned}$$

Hoffman's bound (Lemma 32) applied to this fixed polyhedron gives $\mathrm{dist}_\infty(y, \Omega(M)) \leq \kappa_0 e_N$. It does not require a uniform bound on Hoffman constants of moving constraint matrices.

Transport of the selected optimum to $\Omega(N)$. Using the fixed-chain perturbation identity of Schweitzer [29], Cho and Meyer [8], keep the policy $\pi^\dagger$ fixed and let K_N be its state kernel under N. The common transition support implies that K_N has the same support as $K^\dagger$, so it is irreducible and has a unique stationary row vector d_N. Put $Z = (I - K^\dagger + \mathbf{1}d)^{-1}$. Stationarity and normalization give

$$(d_N - d)(I - K^\dagger) = d_N(K_N - K^\dagger), \qquad (d_N - d)\mathbf{1}d = 0,$$

and therefore

$$d_N - d = d_N(K_N - K^\dagger)Z.$$

Since $\|K_N - K^\dagger\|_{\infty\to\infty} \le e_N$ and $\|d_N\|_1 = 1$, we obtain $\|d_N - d\|_1 \le \kappa_1 e_N$. The allocation $\bar{x}(s,a) = d_N(s)\pi^\dagger(a|s)$ lies in $\Omega(N)$ and satisfies $\|\bar{x} - x\|_\infty \le \kappa_1 e_N$. This is precisely the second transport direction needed in (A3); transporting every true feasible allocation in that direction is unnecessary. □

Pointwise sharpness (A4). At the true model the feasible set is compact and J_M is continuous. Define

$$\varphi_M(u) = \sup\{\mathrm{dist}_\infty(v, O(M)) : v \in \Omega(M),\ J_M^* - J_M(v) \le u\}, \qquad u \ge 0.$$

This is nondecreasing, is zero at zero, and bounds every feasible distance by construction. It also tends to zero as $u \downarrow 0$. Otherwise a sequence of near-maximizers at distance at least some $\epsilon > 0$ would have a convergent subsequence in $\Omega(M)$, and continuity would make its limit a maximizer, a contradiction. This is a pointwise assertion at M and does not give a modulus uniform over nearby models.

Positive state marginals. The target allocation x is stationary for its induced policy, including the arbitrary row convention at zero-mass states. Under Assumption 3, that policy's state kernel is irreducible, and its stationary distribution is strictly positive at every state. Hence $\rho_M > 0$. No uniform positive lower bound on every action probability is imposed.

Lemma 32 (Hoffman error bound)**.** *Let*

$$S = \{x \in \mathbb{R}^n : Ax \le b,\ Cx = d\}$$

be a nonempty polyhedron. Then, for any choice of norms, there exists a constant $\kappa > 0$ such that for all x,

$$\mathrm{dist}(x, S) \ \le\ \kappa\Big(\|(Ax - b)_+\| + \|Cx - d\|\Big).$$

Proof. This is Hoffman's lemma; see [15] and [35, equation 1.3]. □

The following consequence of Hoffman's lemma will also be useful later.

Lemma 33 (LP error bound from Hoffman)**.** *Let $P \subset \mathbb{R}^{d+1}$ be a nonempty polyhedron, and let*

$$t^\star = \max\{t : (\omega, t) \in P\}.$$

Assume the maximum is attained, and define

$$S^\star := P \cap \{t = t^\star\}.$$

Then there exists a constant $\kappa_P > 0$ such that for all $(\omega, t) \in P$,

$$\mathrm{dist}_\infty\big((\omega, t), S^\star\big) \le \kappa_P\,(t^\star - t).$$

Proof. Since P is a polyhedron, there exist matrices A, C and vectors b, d such that

$$P = \{x \in \mathbb{R}^{d+1} : Ax \le b,\ Cx = d\}.$$

Hence

$$S^\star = \{x \in \mathbb{R}^{d+1} : Ax \le b,\ Cx = d,\ t \le t^\star,\ -t \le -t^\star\}$$

is also a nonempty polyhedron. Applying Hoffman's error bound to this system, there exists $\kappa_P > 0$ such that for all $x = (\omega, t)$,

$$\mathrm{dist}_\infty(x, S^\star) \le \kappa_P\Big(\|(Ax - b)_+\|_\infty + \|Cx - d\|_\infty + (t - t^\star)_+ + (t^\star - t)_+\Big).$$

Now let $x = (\omega, t) \in P$. Then $(Ax - b)_+ = 0$ and $Cx - d = 0$. Moreover, since $t^\star$ is the optimal value over P, we have $t \le t^\star$, so $(t - t^\star)_+ = 0$. Therefore

$$\mathrm{dist}_\infty\big((\omega, t), S^\star\big) \le \kappa_P\,(t^\star - t),$$

as claimed. □

D.3 Linear Sharpness Under a Finite Discretisation of the MDP Space

In many settings it is natural to approximate the alternative family by a finite discretization. For example, if the model class is parametrized by a compact parameter space, one may construct a finite grid approximation of the alternatives. Such discretizations are also common in algorithmic implementations and in lower bound arguments, where a finite collection of alternative models suffices to capture the intrinsic difficulty of the identification problem.

In this section we show that a finite alternative objective has linear sharpness. This does not justify replacing a continuum of alternatives by a finite grid in a correctness certificate: the resulting larger information objective need not represent the original identification problem. The key observation is that once the set of alternatives is finite, the function $J_M(\omega)$ becomes a concave piecewise-linear function of ω. Maximizing such a function over the polyhedral feasible set $\Omega(M)$ can be written as a linear program in hypograph form. Classical error bounds for linear programs then imply a linear relationship between the suboptimality gap and the distance to the optimal set.

We write this into a lemma and proof below. Informally, when the alternative family is discretized, the objective becomes piecewise linear and the optimization problem reduces to a linear program. In this case the suboptimality gap grows linearly with the distance to the optimal set, yielding a linear pointwise modulus at M.

Discretized objective. Let $\mathrm{Alt}_\Delta(M) = \{M^1, \dots, M^N\} \subset \mathrm{Alt}(M)$ be a finite subset of alternatives. Define the discretized objective

$$J_\Delta(M, \omega) := \min_{1 \le i \le N} \langle \omega, k(M^i) \rangle,$$

where $k(M^i)$ is the vector with entries $k(M^i)(s,a) = \mathrm{KL}(P_M(\cdot|s,a), P_{M^i}(\cdot|s,a))$. Let

$$J^\star_\Delta(M) := \max_{\omega \in \Omega(M)} J_\Delta(M, \omega), \qquad S_\Delta(M) := \arg\max_{\omega \in \Omega(M)} J_\Delta(M, \omega).$$

Proposition 34 (Linear sharpness under discretization)**.** *Assume that $\Omega(M) \subset \mathbb{R}^d$ is a nonempty compact polytope. Let*

$$\mathrm{Alt}_\Delta(M) = \{M_1, \dots, M_N\} \subset \mathrm{Alt}(M)$$

be a finite discretization of the alternative set, and define

$$J_\Delta(M, \omega) := \min_{1 \le i \le N} \langle \omega, k(M_i) \rangle, \qquad \omega \in \Omega(M).$$

Let

$$J^\star_\Delta(M) := \max_{\omega \in \Omega(M)} J_\Delta(M, \omega), \qquad S_\Delta(M) := \arg\max_{\omega \in \Omega(M)} J_\Delta(M, \omega).$$

Then there exists a constant $c_\Delta(M) > 0$ such that for all $\omega \in \Omega(M)$,

$$\mathrm{dist}_\infty\big(\omega, S_\Delta(M)\big) \le c_\Delta(M) \Big(J^\star_\Delta(M) - J_\Delta(M, \omega) \Big).$$

Consequently, the pointwise condition (A4) in Lemma 30 holds for the discretized objective with linear modulus

$$\varphi_M(u) = c_\Delta(M)\, u.$$

Proof. Write $k_i := k(M_i) \in \mathbb{R}^d$ for $i = 1, \dots, N$. Since the alternative set is finite,

$$J_\Delta(M, \omega) = \min_{1 \le i \le N} \langle \omega, k_i \rangle$$

is a concave piecewise-linear function of ω. Now, introduce the lifted feasible set

$$P_\Delta(M) := \Big\{ (\omega, t) \in \mathbb{R}^{d+1} : \omega \in \Omega(M),\ t \le \langle \omega, k_i \rangle \ \forall i = 1, \dots, N \Big\}.$$

Then maximizing $J_\Delta(M, \omega)$ over $\Omega(M)$ is equivalent to the linear program

$$J^\star_\Delta(M) = \max\{ t : (\omega, t) \in P_\Delta(M) \}.$$

Because $\Omega(M)$ is a polytope and the additional constraints $t \le \langle \omega, k_i \rangle$ are linear, $P_\Delta(M)$ is a nonempty polyhedron. Since $\Omega(M)$ is compact and each map $\omega \mapsto \langle \omega, k_i \rangle$ is continuous, the

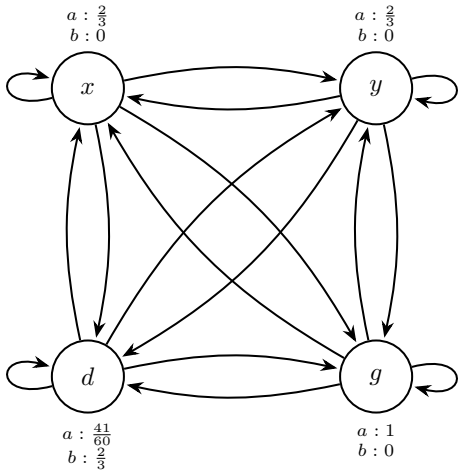

Figure 2: Visualization of the family of MDPs used in the construction in Proposition 35. We display the deterministic rewards from the two actions a, b next to each of the states. The MDP is fully connected and the transitions vary parametrically through the family of MDPs. See Proposition 35 for more details.

function $J_\Delta(M, \omega) = \min_i \langle \omega, k_i \rangle$ is continuous on $\Omega(M)$. Hence the maximum $J^\star_\Delta(M)$ is attained. Define the lifted optimal set

$$S^{\text{lift}}_\Delta(M) := P_\Delta(M) \cap \{t = J^\star_\Delta(M)\}.$$

By Lemma 33, there exists a constant $\kappa_\Delta(M) > 0$ such that for all $(\omega, t) \in P_\Delta(M)$,

$$\operatorname{dist}_\infty\big((\omega, t), S^{\text{lift}}_\Delta(M)\big) \le \kappa_\Delta(M)\big(J^\star_\Delta(M) - t\big).$$

Now fix $\omega \in \Omega(M)$, and set $t := J_\Delta(M, \omega)$. By definition of J_Δ, we have $t \le \langle \omega, k_i \rangle$ for all i, hence $(\omega, t) \in P_\Delta(M)$. Therefore

$$\operatorname{dist}_\infty\big((\omega, J_\Delta(M, \omega)), S^{\text{lift}}_\Delta(M)\big) \le \kappa_\Delta(M)\Big(J^\star_\Delta(M) - J_\Delta(M, \omega)\Big).$$

Finally, projection onto the ω-coordinates cannot increase ℓ_∞-distance, so

$$\operatorname{dist}_\infty\big(\omega, S_\Delta(M)\big) \le \kappa_\Delta(M)\Big(J^\star_\Delta(M) - J_\Delta(M, \omega)\Big).$$

Setting $c_\Delta(M) := \kappa_\Delta(M)$ completes the proof. □

D.4 An explicit family realizing arbitrary rational exponents for polynomial sharpness

We construct a structured family with sharpness-modulus exponent $(m - n)/m$ for any integers $m > n \ge 1$. These integers are local to this example and are unrelated to the connectivity parameter in the coverage bound. The objective depends only on the allocation to (d, b), which is also the distance to the full optimal set. Thus the directional expansion below gives a modulus for every feasible allocation, as verified in Step 10.

Proposition 35 (Explicit communicating and aperiodic family realizing every rational exponent > 1)**.** *Fix integers $m > n \ge 1$. Then there exist:*

- *a finite discounted communicating and aperiodic MDP M with $|\mathcal{S}| = 4$, $|\mathcal{A}| = 2$, rewards in $[0, 1]$, and unique deterministic optimal policy $\pi^\star$,*
- *a compact semialgebraic feasible family Θ of communicating and aperiodic models,*
- *a known reference kernel μ such that $P_M \in G_\mu(L)$ for every $L > 1$, and in fact $\Theta \subseteq G_\mu(4)$ with μ uniform,*

- *a boundary point* $\omega^{\mathrm{ref}} \in \Omega^\star(M)$*,*
- *and a feasible transverse direction* h*, i.e., such that* $\omega^{\mathrm{ref}} + th \in \Omega(M)$ *for all sufficiently small* $t > 0$

such that

$$J^\star - J_M(\omega^{\mathrm{ref}} + th) = c_{m,n}\, t^{m/(m-n)} + o\big(t^{m/(m-n)}\big) \qquad (t \downarrow 0), \tag{52}$$

where

$$c_{m,n} = \frac{4}{3}\Big(1 - \frac{n}{m}\Big)\Big(\frac{4n}{m}\Big)^{\frac{n}{m-n}} > 0.$$

In particular, every rational exponent $p > 1$ *is achievable even within the communicating and aperiodic discounted setting.*

Proof. We divide the proof into several steps.

Step 1: the true MDP and the reward normalization. Let

$$\mathcal{S} = \{x, y, d, g\}, \qquad \mathcal{A} = \{a, b\}, \qquad \gamma = \frac{1}{2}.$$

Define the true transition kernel by

$$P_M(\cdot \mid s, a) = P_M(\cdot \mid s, b) = p \qquad \forall s \in \mathcal{S},$$

where

$$p := \Big(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\Big)$$

in the state order (x, y, d, g). Thus every row of every transition matrix equals p. In particular, under every deterministic policy the induced state chain has strictly positive transition probabilities, so M is communicating and aperiodic.

The actual reward function used in the proposition is

$$r(x, a) = r(y, a) = \frac{2}{3}, \qquad r(d, a) = \frac{41}{60}, \qquad r(g, a) = 1,$$

and

$$r(x, b) = r(y, b) = r(g, b) = 0, \qquad r(d, b) = \frac{2}{3}.$$

Hence $r(s, \alpha) \in [0, 1]$ for every state-action pair. For algebraic convenience, introduce the auxiliary reward

$$\bar{r} := 3r - 2,$$

so that

$$\bar{r}(x, a) = \bar{r}(y, a) = 0, \qquad \bar{r}(d, a) = \frac{1}{20}, \qquad \bar{r}(g, a) = 1,$$

and

$$\bar{r}(x, b) = \bar{r}(y, b) = \bar{r}(g, b) = -2, \qquad \bar{r}(d, b) = 0.$$

The deterministic optimal-policy sets under r and $\bar{r}$ coincide for the true model and for every alternative considered below. Therefore all optimality calculations may be carried out with $\bar{r}$.

Under M, the transition kernel is independent of the action, so action comparisons are determined entirely by immediate rewards. Since

$$\bar{r}(s, a) - \bar{r}(s, b) > 0 \qquad \forall s \in \mathcal{S},$$

it follows that the policy

$$\pi^\star(s) \equiv a \qquad (s \in \mathcal{S})$$

is the unique deterministic optimal policy of M. The same is true for the actual reward function $r \in [0, 1]$.

Step 2: the alternative family and the $G_\mu(L)$ condition. Fix $\bar{u} \in (0, 1/16]$. For $u \in [0, \bar{u}]$, define two probability vectors

$$q_a(u) := \Big(\frac{3}{8} + u^m, \frac{1}{4}, \frac{1}{4}, \frac{1}{8} - u^m\Big), \qquad q_b(u) := \Big(\frac{1}{8} + u^n, \frac{1}{4}, \frac{1}{4}, \frac{3}{8} - u^n\Big),$$

again in the state order (x, y, d, g). Because $\bar{u} \le 1/16$, all entries of $q_a(u)$ and $q_b(u)$ belong to $[1/16, 7/16]$, in particular they are strictly positive.

Now define M_u by modifying only the two transitions at state d:

$$P_{M_u}(\cdot \mid d, a) = q_a(u), \qquad P_{M_u}(\cdot \mid d, b) = q_b(u),$$

while for every other state-action pair $(s, \alpha) \notin \{(d, a), (d, b)\}$,

$$P_{M_u}(\cdot \mid s, \alpha) = p.$$

Set

$$\Theta := \{M\} \cup \{M_u : 0 \le u \le \bar{u}\}.$$

The family Θ is compact, and because it is parameterized polynomially by u, it is semialgebraic as a subset of the product of simplices. Since every transition probability of every M_u is strictly positive, every M_u is also communicating and aperiodic.

Define the reference kernel μ by

$$\mu(\cdot \mid s, \alpha) = p \qquad \forall (s, \alpha) \in \mathcal{S} \times \mathcal{A}.$$

Then $P_M = \mu$, so $P_M \in G_\mu(L)$ for every $L > 1$. Moreover, because every entry of every transition row in the family belongs to $[1/16, 7/16]$, one has for every $M_u \in \Theta$ and every state-action-state triple (s, α, s'),

$$\frac{\mu(s' \mid s, \alpha)}{P_{M_u}(s' \mid s, \alpha)} \in \left[\frac{1/4}{7/16}, \frac{1/4}{1/16}\right] = \left[\frac{4}{7}, 4\right] \subseteq [1/4, 4].$$

Hence $\Theta \subseteq G_\mu(4)$.

Step 3: the alternative policy is strictly optimal for all small u. Define the deterministic policy

$$\pi_u(x) = a, \qquad \pi_u(y) = a, \qquad \pi_u(d) = b, \qquad \pi_u(g) = a.$$

We first verify strict optimality at $u = 0$. Under M_0, the policy π_0 has value vector

$$V_{M_0}^{\pi_0} = \Big(\frac{17}{64}, \frac{17}{64}, \frac{21}{64}, \frac{81}{64}\Big),$$

which is obtained by solving the Bellman system

$$\begin{aligned}
V(x) &= \frac{1}{2} \cdot \frac{V(x) + V(y) + V(d) + V(g)}{4}, \\
V(y) &= \frac{1}{2} \cdot \frac{V(x) + V(y) + V(d) + V(g)}{4}, \\
V(d) &= \frac{1}{2}\Big(\frac{1}{8}V(x) + \frac{1}{4}V(y) + \frac{1}{4}V(d) + \frac{3}{8}V(g)\Big), \\
V(g) &= 1 + \frac{1}{2} \cdot \frac{V(x) + V(y) + V(d) + V(g)}{4}.
\end{aligned}$$

Now compare actions state by state.

At x and y, action b has the same transition kernel as action a but a reward smaller by 2, so

$$Q_{M_0}^{\pi_0}(x, a) - Q_{M_0}^{\pi_0}(x, b) = 2, \qquad Q_{M_0}^{\pi_0}(y, a) - Q_{M_0}^{\pi_0}(y, b) = 2.$$

At g, action b has the same transition kernel as a but a reward smaller by 3, so

$$Q_{M_0}^{\pi_0}(g, a) - Q_{M_0}^{\pi_0}(g, b) = 3.$$

At d, one computes

$$Q_{M_0}^{\pi_0}(d, b) = \frac{1}{2}\Big(\frac{1}{8} \cdot \frac{17}{64} + \frac{1}{4} \cdot \frac{17}{64} + \frac{1}{4} \cdot \frac{21}{64} + \frac{3}{8} \cdot \frac{81}{64}\Big) = \frac{21}{64},$$

whereas

$$Q_{M_0}^{\pi_0}(d, a) = \frac{1}{20} + \frac{1}{2}\Big(\frac{3}{8} \cdot \frac{17}{64} + \frac{1}{4} \cdot \frac{17}{64} + \frac{1}{4} \cdot \frac{21}{64} + \frac{1}{8} \cdot \frac{81}{64}\Big) = \frac{81}{320}.$$

Hence

$$Q^{\pi_0}_{M_0}(d,b) - Q^{\pi_0}_{M_0}(d,a) = \frac{3}{40} > 0.$$

Thus π_0 is the unique deterministic optimal policy of M_0.

In fact, no further shrinking of $\bar{u} \le 1/16$ is needed. Under π_u, the identical transition rows at x and g imply $V(g) - V(x) = 1$ for the auxiliary reward. The action gap at d is therefore

$$Q^{\pi_u}_{M_u}(d,b) - Q^{\pi_u}_{M_u}(d,a) = -\frac{1}{20} + \frac{1}{2}\left(\frac{1}{4} - u^n + u^m\right) = \frac{3}{40} - \frac{1}{2}u^n + \frac{1}{2}u^m \ge \frac{3}{40} - \frac{1}{32} > 0.$$

The other three gaps remain $2, 2, 3$. Hence π_u is uniquely optimal for every $u \in [0, \bar{u}]$, both under $\bar{r}$ and the actual reward r. With the admissible class chosen to be exactly $\mathcal{M} = \Theta$, it follows that

$$\mathrm{Alt}(M) = \{M_u : 0 \le u \le \bar{u}\}.$$

The choice of this structured class is essential to the objective below; allowing all of $G_\mu(4)$ would give a different alternative set.

Step 4: the navigation polytope. Because every row of P_M equals p, the balance equations become

$$\omega(s) = \sum_{s',\alpha} p(s)\,\omega(s',\alpha) = p(s)\sum_{s',\alpha}\omega(s',\alpha) = p(s) \qquad \forall s \in \mathcal{S}.$$

Since $\omega \in \Delta(\mathcal{S} \times \mathcal{A})$, this is equivalent to

$$\omega(x) = \omega(y) = \omega(d) = \omega(g) = \frac{1}{4}.$$

Hence

$$\Omega(M) = \Big\{\omega \in \Delta(\mathcal{S} \times \mathcal{A}) : \omega(x) = \omega(y) = \omega(d) = \omega(g) = \frac{1}{4}\Big\}.$$

Step 5: the objective and the optimal-allocation set. For $u \in [0, \bar{u}]$, define

$$K_m(u) := \mathrm{KL}\big(p \,\|\, q_a(u)\big), \qquad L_n(u) := \mathrm{KL}\big(p \,\|\, q_b(u)\big).$$

Since every state-action pair other than (d,a) and (d,b) has the same transition under M and M_u, one has, for every $\omega \in \Omega(M)$,

$$J_M(\omega) = \inf_{u \in [0,\bar{u}]}\Big[\omega(d,a)K_m(u) + \omega(d,b)L_n(u)\Big].$$

Because $\omega(d,a) + \omega(d,b) = 1/4$ for every $\omega \in \Omega(M)$, this is a one-dimensional action-allocation problem at state d.

A direct computation gives

$$K_m(0) = L_n(0) =: A = \frac{1}{4}\log\frac{4}{3}.$$

Moreover, for $u > 0$,

$$K_m(u) > A > L_n(u).$$

If $\omega(d,b) = 0$, then $\omega(d,a) = 1/4$ and

$$J_M(\omega) = \inf_{u \in [0,\bar{u}]} \frac{1}{4}K_m(u) = \frac{1}{4}A.$$

If $\omega(d,b) > 0$, then choosing $u > 0$ small enough yields

$$\omega(d,a)K_m(u) + \omega(d,b)L_n(u) < \frac{1}{4}A,$$

because the negative u^n variation in $L_n(u)$ dominates the positive u^m variation in $K_m(u)$ as $u \downarrow 0$. Hence

$$J_M(\omega) < \frac{1}{4}A \qquad \text{whenever } \omega(d,b) > 0.$$

Therefore

$$J^\star = \frac{1}{4}A$$

and

$$\Omega^\star(M) = \{\omega \in \Omega(M) : \omega(d,b) = 0,\ \omega(d,a) = 1/4\}.$$

Choose

$$\omega^{\mathrm{ref}}(s,a) = \frac{1}{4}, \qquad \omega^{\mathrm{ref}}(s,b) = 0 \qquad (s \in \mathcal{S}).$$

Then $\omega^{\mathrm{ref}} \in \Omega^\star(M)$ is a boundary reference point. It is not the minimum-norm selector $x = \omega^*(M)$ used by the algorithm: that selector has $x(d,a) = 1/4$, $x(d,b) = 0$, and $x(s,a) = x(s,b) = 1/8$ for $s \neq d$. Both induce irreducible chains, since every transition row is positive.

Step 6: the transverse direction. Let

$$h := e_{(d,b)} - e_{(d,a)}.$$

Because h preserves the state marginal at d and leaves all other state marginals unchanged, it is a feasible direction at ω^{ref}. It is not strongly tangent to $\Omega^\star(M)$, since

$$(\omega^{\mathrm{ref}} + th)(d,b) = t > 0 \qquad (t > 0),$$

whereas every point of $\Omega^\star(M)$ satisfies $\omega(d,b) = 0$. Thus h is transverse.

For $t \in (0, 1/4)$, write

$$\omega_t := \omega^{\mathrm{ref}} + th.$$

Then

$$\omega_t(d,a) = \frac{1}{4} - t, \qquad \omega_t(d,b) = t,$$

and therefore

$$J_M(\omega_t) = \inf_{u \in [0,\bar{u}]} F_t(u), \qquad F_t(u) := \Big(\frac{1}{4} - t\Big)K_m(u) + tL_n(u).$$

Note also that the infinity norm distance between ω_t and $\Omega^*(M)$ is t in this case. Namely $\mathrm{dist}_\infty(\omega_t, \Omega^*(M)) = t$. Hence, the final bounds on the curvature can be related to our definition of sharpness in Assumption (A4) of Lemma 30.

Step 7: asymptotic expansions of the branch costs. Because

$$q_a(u) = \Big(\frac{3}{8} + u^m, \frac{1}{4}, \frac{1}{4}, \frac{1}{8} - u^m\Big), \qquad q_b(u) = \Big(\frac{1}{8} + u^n, \frac{1}{4}, \frac{1}{4}, \frac{3}{8} - u^n\Big),$$

one computes

$$K_m(u) = \frac{1}{4}\log\frac{(1/4)^2}{(3/8 + u^m)(1/8 - u^m)}, \qquad L_n(u) = \frac{1}{4}\log\frac{(1/4)^2}{(1/8 + u^n)(3/8 - u^n)}.$$

Expanding at $u = 0$ gives

$$K_m(u) = A + \frac{4}{3}u^m + O(u^{2m}), \qquad L_n(u) = A - \frac{4}{3}u^n + O(u^{2n}).$$

Hence

$$F_t(u) = \frac{A}{4} + \frac{1}{3}u^m - \frac{4}{3}tu^n + R_t(u),$$

where

$$|R_t(u)| \le C\big(tu^m + u^{2m} + tu^{2n}\big)$$

for all sufficiently small u, t.

Step 8: localization of the minimizer. Set $s = 1/(m-n)$ and $p^\sharp = m/(m-n)$. The continuous function F_t attains its minimum on $[0, \bar{u}]$. Moreover $F_t \to K_m/4$ uniformly as $t \downarrow 0$, and $K_m(u) > K_m(0)$ for $u > 0$. Every minimizing sequence u_t therefore tends to zero: otherwise a subsequential limit away from zero would minimize K_m, a contradiction.

We strengthen this to $u_t = O(t^s)$. Choose $M_0 > 0$ with $M_0^{m-n} \geq 8$. If $u \geq M_0 t^s$, then

$$\frac{1}{3}u^m - \frac{4}{3}tu^n \geq \frac{1}{6}u^m.$$

In this range the remainder satisfies

$$\frac{|R_t(u)|}{u^m} \leq C\left(t + u^m + tu^{2n-m}\right) \leq C\left(t + u^m + \frac{u^n}{M_0^{m-n}}\right).$$

If a minimizer satisfied this inequality on u, its right-hand side would tend to zero because $u_t \to 0$. Consequently $F_t(u_t) > A/4$ for sufficiently small t. But the feasible trial point $u = t^s$ has

$$F_t(t^s) = A/4 + t^{p^\sharp}(-1 + o(1)) < A/4.$$

This contradiction proves $u_t < M_0 t^s$ eventually, uniformly over choices of minimizer.

Step 9: scaling and the sharpness exponent. Set $u = t^s v$. Since $u_t = O(t^s)$, it suffices to analyze bounded v. Uniformly for v in bounded sets,

$$F_t(t^s v) = \frac{A}{4} + t^{p^\sharp}\Big[\frac{1}{3}v^m - \frac{4}{3}v^n + o(1)\Big].$$

Define

$$\psi(v) := \frac{1}{3}v^m - \frac{4}{3}v^n.$$

On the relevant domain $v \geq 0$, this polynomial has a unique minimizer at

$$v_\star = \Big(\frac{4n}{m}\Big)^{1/(m-n)},$$

and

$$\psi(v_\star) = -\frac{4}{3}\Big(1 - \frac{n}{m}\Big)\Big(\frac{4n}{m}\Big)^{\frac{n}{m-n}} =: -c_{m,n} < 0.$$

Therefore,

$$\inf_{u \in [0,\bar{u}]} F_t(u) = \frac{A}{4} - c_{m,n}\, t^{p^\sharp} + o(t^{p^\sharp}).$$

Since $J^\star = A/4$, this gives

$$J^\star - J_M(\omega^{\mathrm{ref}} + th) = c_{m,n}\, t^{m/(m-n)} + o\big(t^{m/(m-n)}\big),$$

where

$$c_{m,n} = \frac{4}{3}\Big(1 - \frac{n}{m}\Big)\Big(\frac{4n}{m}\Big)^{\frac{n}{m-n}} > 0.$$

Step 10: a sharpness modulus for the full optimal set. For every $\omega \in \Omega(M)$, put $t_\omega = \omega(d, b) \in [0, 1/4]$. Step 5 gives

$$J_M(\omega) = I(t_\omega), \qquad I(t) = \inf_{u \in [0,\bar{u}]}\big[(1/4 - t)K_m(u) + tL_n(u)\big].$$

Every optimal allocation has zero mass at (d, b), so its distance from ω is at least t_ω. Moving precisely this mass from (d, b) to (d, a) preserves all state marginals and gives an optimal allocation. Hence

$$\mathrm{dist}_\infty(\omega, \Omega^*(M)) = t_\omega.$$

Let $p = (m-n)/m \in (0, 1)$. Step 9 shows $J_M^* - I(t) = c_{m,n} t^{1/p}(1 + o(1))$ as $t \downarrow 0$, independently of the other coordinates of ω. Consequently, for all sufficiently small t_ω,

$$\mathrm{dist}_\infty(\omega, \Omega^*(M)) \leq (2/c_{m,n})^p [J_M^* - J_M(\omega)]^p.$$

The continuous gap $J_M^* - I(t)$ is strictly positive on every compact subinterval of $(0, 1/4]$ by Step 5. Increasing the constant therefore makes this inequality valid on all of $\Omega(M)$. Conversely, the same expansion rules out any modulus of order $u^{p'}$ with $p' > p$. This proves the sharpness statement in Theorem 4 over the whole feasible allocation set.

□

D.5 Empirical Verification of Sharpness Results.

Construction of Figure 1. For each pair of integers (m, n) shown in the figure, we instantiate the MDP family from Theorem 4. Let $\Omega^\star(M)$ denote the set of optimal allocations and let $\omega^{\mathrm{ref}} \in \Omega^\star(M)$ be the reference optimal allocation used in the construction. We then consider the one-dimensional transverse path

$$\omega_t = \omega^{\mathrm{ref}} + t\big(e_{(d,b)} - e_{(d,a)}\big),$$

which moves mass from the optimal state-action pair (d, a) to the suboptimal state-action pair (d, b) while remaining feasible for sufficiently small $t > 0$. As described in Appendix D.4, for each value of t, we compute the objective value

$$J(\omega_t) = \inf_{u \in [0,1/16]} \sum_{s,a} \omega_t(s,a) \mathrm{KL}(P(\cdot \mid s,a) \,\|\, P_u(\cdot \mid s,a)),$$

where P_u ranges over the one-parameter family of alternatives used in the proof. The scalar minimization over u is solved numerically using the reduced one-dimensional objective from the construction. We then plot the resulting gaps $J^\star - J(\omega_t)$ against t on log-log axes and fit a line to each curve in log space. The fitted slope estimates the exponent in

$$J^\star - J(\omega_t) \asymp t^{m/(m-n)},$$

so the reciprocal fitted slope estimates the polynomial sharpness exponent of φ, equal to $(m-n)/m$.

The experiments were conducted on a cluster equipped with NVIDIA RTX 6000 Ada Generation GPUs (48 GB VRAM each).

E Other results

Under the reward condition below, the set of transition kernels with multiple optimal policies has Lebesgue measure zero in the affine coordinates of the full product of transition simplices. This statement does not by itself apply to relative Lebesgue measure on an arbitrary fixed-support face or structured family, where a tie equation can vanish identically. To prove this result, we employ the following lemma.

Lemma 36 (Rational structure of policy values)**.** *Consider the set of tabular MDPs with fixed reward vector $r \in [0,1]^{\mathcal{S}\times\mathcal{A}}$ and $\gamma \in (0,1)$. Let M be an MDP identified by $P \in \Delta(\mathcal{S})^{\mathcal{S}\times\mathcal{A}}$. For any two policies (π, π'), for any $s \in \mathcal{S}$ the map*

$$g_M^{\pi,\pi'}(s) \coloneqq V_M^{\pi}(s) - V_M^{\pi'}(s),$$

is a real rational function.

Proof. Since $V_M^{\pi} = (I-\gamma P_\pi)^{-1} r_\pi$, by Cramer's rule, $(I-\gamma P_\pi)^{-1} = \mathrm{adj}(I-\gamma P_\pi)/\det(I-\gamma P_\pi)$, with $\det(I-\gamma P_\pi) \neq 0$ since $\gamma \in (0,1)$. Entries of $\mathrm{adj}(\cdot)$ and the determinant are polynomials in the entries of P_π, hence V_M^{π} is componentwise a ratio of polynomials in (P, r) for every s. □

We also need the following useful result, showing that optimal policies do not yield constant reward vectors if $\max_a r(s,a)$ varies across states.

Lemma 37 (Constant-reward policies cannot be optimal)**.** *Let $m(s) := \max_a r(s,a)$ and assume $\exists s, s' : m(s) \neq m(s')$. If a deterministic policy π has $r_\pi(s) \equiv c$ for all $s \in \mathcal{S}$, then π is not optimal for any transition kernel P.*

Proof. If $r_\pi = c\mathbf{1}$ for some $c \in [0,1]$, then for any P we have $V_M^{\pi} = \frac{c}{1-\gamma}\mathbf{1}$. For any s and b,

$$Q_M^{\pi}(s,b) = r(s,b) + \gamma\frac{c}{1-\gamma}, \qquad V_M^{\pi}(s) = \frac{c}{1-\gamma}.$$

Optimality of π would require $V_M^{\pi}(s) \geq Q_M^{\pi}(s,b)$ for all (s,b), i.e.

$$0 \leq \frac{c}{1-\gamma} - r(s,b) - \gamma\frac{c}{1-\gamma} = c - r(s,b) \Rightarrow c \geq r(s,b)$$

for all (s,b), implying $c \geq m(s)$ for all s. But $c = r_\pi(s) = r(s,\pi(s)) \leq m(s)$ for all s, implying we must have $c = m(s)$ for all s, which is however a contradiction since there exist a pair (s,s') s.t. $m(s) \neq m(s')$. □

Next, we show that g^π is non-trivial under some assumptions.

Lemma 38 (Non-triviality of g)**.** *Fix deterministic π, a state s, and $a \neq \pi(s)$. Define*

$$g_{s,a}^{\pi}(M) := Q_M^{\pi}(s,\pi(s)) - Q_M^{\pi}(s,a).$$

Then $g_{s,a}^{\pi} \equiv 0$ on $\Delta(\mathcal{S})^{\mathcal{S}\times\mathcal{A}}$ if and only if $r(s,\pi(s)) = r(s,a)$ and r_π is constant across states.

Proof. Write

$$g_{s,a}^{\pi}(M) = [r(s,\pi(s)) - r(s,a)] + \gamma(p-q)^\top V^{\pi}(P)$$

for some MDP with transition P and we write $p = P(\cdot|s,\pi(s))$, $q = P(\cdot|s,a)$.

Sufficiency. If $r(s,\pi(s)) = r(s,a)$ and $r_\pi = c\mathbf{1}$ for some $c \in [0,1]$, then

$$g_{s,a}^{\pi}(M) = c\gamma(p-q)^\top(I-\gamma P_\pi)^{-1}\mathbf{1}$$

Since $\mathbf{1}$ is a right eigenvector of $(I-\gamma P_\pi)^{-1}$ satisfying $(I-\gamma P_\pi)^{-1}\mathbf{1} = \frac{1}{1-\gamma}\mathbf{1}$, we have that

$$g_{s,a}^{\pi}(M) = \frac{c\gamma}{1-\gamma}(p-q)^\top\mathbf{1} = 0.$$

which proves sufficiency.

Necessity. Suppose $g^{\pi}_{s,a}(M) \equiv 0$ on $P \in \Delta(\mathcal{S})^{\mathcal{S}\times\mathcal{A}}$. For each P we obtain P_π. Treat P_π as a fixed variable, while q is a degree of freedom. Then, we must have that

$$0 = \sup_q \left| [r(s,\pi(s)) - r(s,a)] + \gamma(p-q)^\top V^\pi(P) \right|.$$

Since for any q we obtain 0, consider $q = e_i$ and $q = e_j$ for i, j taking value in an enumeration of $\mathcal{S}$. Then

$$[r(s,\pi(s)) - r(s,a)] + \gamma(p-e_i)^\top V^\pi(P) = [r(s,\pi(s)) - r(s,a)] + \gamma(p-e_j)^\top V^\pi(P)$$

implying that $(e_j - e_i)^\top V^\pi(P) = 0$. Since i, j are allowed to vary, this implies that $V^\pi_P = c\mathbf{1}$ for some $c \in \mathbb{R}$. Therefore

$$g^{\pi}_{s,a}(M) = [r(s,\pi(s)) - r(s,a)] + \gamma(p-q)^\top c\mathbf{1} = r(s,\pi(s)) - r(s,a),$$

implying that $r(s,\pi(s)) = r(s,a)$ if $g^{\pi}_{s,a}(M) = 0$.

Now, for any fixed q consider varying P so that P_π varies. Since $V^\pi_P = c\mathbf{1} = (I - \gamma P_\pi)^{-1} r_\pi$, and $\mathbf{1}$ is the right eigenvector of $(I - \gamma P_\pi)^{-1}$, we have that r_π itself is proportional to $\mathbf{1}$. Thus we conclude that r_π must be constant across states. □

Lemma 39 (Switch hypersurfaces have null measure)**.** *Let* $m(s) = \max_a r(s,a)$, $Z^{\pi}_{s,a} \coloneqq \{P \in \Delta(\mathcal{S})^{\mathcal{S}\times\mathcal{A}} : g^{\pi}_{s,a}(P) = 0\}$ *and* $\mathcal{U}^\pi \coloneqq \{P \in \Delta(\mathcal{S})^{\mathcal{S}\times\mathcal{A}} : \pi \text{ is optimal at } (P, r, \gamma)\}$.

If $\exists s, s' : m(s) \neq m(s')$ *, then for every deterministic* π *and every* $a \neq \pi(s)$*, the set* $\mathcal{U}^\pi \cap Z^{\pi}_{s,a}$ *has Lebesgue measure* 0*.*

Proof. By the previous lemma we know that $g^{\pi}_{s,a}$ is a real rational function, thus real analytic. In affine simplex coordinates, a nontrivial such function has a nonzero polynomial numerator. Its zero set has Lebesgue measure zero: this follows by induction on dimension, using finitely many roots in one dimension and Fubini after excluding the common zero set of the coefficient polynomials. The boundary of the product of simplices also has measure zero in these coordinates.

If $g^{\pi}_{s,a} \equiv 0$, then by the previous lemma we have that r_π is constant. But optimal policies do not admit constant r_π if there exists a pair (s, s') such that $m(s) \neq m(s')$, so $\mathcal{U}^\pi = \varnothing$ and $\mathcal{U}^\pi \cap Z^{\pi}_{s,a} = \varnothing$ (null). Alternatively, if $g^{\pi}_{s,a} \not\equiv 0$, then $Z^{\pi}_{s,a}$ is the zero set of a nontrivial real-analytic function, hence it has Lebesgue measure 0; intersecting with the measurable set $\mathcal{U}^\pi$ preserves nullity. □

Lemma 40 (Policy-switch set has Lebesgue measure 0)**.** *Let* $m(s) = \max_a r(s,a)$*. If* $\exists s, s' : m(s) \neq m(s')$ *then the set*

$$\mathcal{T} := \{P \in \Delta(\mathcal{S})^{\mathcal{S}\times\mathcal{A}} : \exists s,\ a \neq b \text{ with } Q^\star_M(s,a) = Q^\star_M(s,b) = V^\star_M(s)\}$$

has Lebesgue measure 0*.*

Proof. Optimal policies are w.l.o.g. deterministic. If $P \in \mathcal{T}$, there exists an optimal deterministic $\pi^\star$, a state s, and $a \neq \pi^\star(s)$ with $Q^{\pi^\star}_M(s,\pi^\star(s)) = Q^{\pi^\star}_M(s,a) = V^\star_M(s)$, hence $P \in \mathcal{U}^{\pi^\star} \cap Z^{\pi^\star}_{s,a}$. Therefore

$$\mathcal{T} \subseteq \bigcup_\pi \bigcup_s \bigcup_{a\neq\pi(s)} \left(\mathcal{U}^\pi \cap Z^{\pi}_{s,a}\right).$$

Since there are finitely many deterministic policies, states, and actions, this is a finite union of measure-zero sets, and we have that $\mathcal{T}$ has measure 0. □

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: Yes, the main claim is the development of non-asymptotic bounds for NaS and the tools to do so. This is stated clearly in the main body, with complementary analysis and details provided in the appendix.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We provide a list of Assumption in Section 2.2 required for our analysis to hold. Furthermore, in the final section we discuss the computational limitations of NaS-style algorithms.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: Detailed proofs are provided in the appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: Any experimental results are discussed in the main paper, with a more detailed discussion in the appendix. Code for reproducibility is provided in the supplementary material.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: Yes, any code necessary is provided in the supplementary material.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Any hyperparameters used in the experiments are discussed.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [N/A]

   Justification: Experiments presented in this work are deterministic and designed for understanding the structure of various important functions. Hence, error bars are not appropriate for code ran and figures shown.

   Guidelines:

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

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Yes, this is discussed in the appendix.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research provided in this paper meets the desired ethical standards.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This work is foundational, with a focus on theoretical guarantees for particular classes of algorithms. While we hope this work provides a deeper understanding of certain RL problems, we believe any societal impacts would be indirect.

    Guidelines:

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

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

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

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

13. **New assets**

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: [N/A]

Guidelines:

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