Adaptive Joint Testing of Policies in Discounted Markov Decision Processes

Po-An Wang
Institute of Statistics and Data Science
National Tsing Hua University
po-an@stat.nthu.edu.tw

Kaito Ariu
CyberAgent
Tohoku University
kaito_ariu@cyberagent.co.jp

Abstract

We study fixed-confidence joint policy testing in discounted tabular Markov decision processes. From one adaptive trajectory, a learner must certify the signs of a fixed finite family of policy values under an unknown transition kernel. We identify an instance-specific characteristic time $T^\star(p)$ through a max–min information program over stationary occupancies and sign-flipped alternatives, and prove the corresponding asymptotic lower bound. We then develop PT-ACE(μ), which updates a shared sampling allocation by the exchange method, selects actions using averaged occupancies, and applies certified stopping tests to each target policy. Under nontriviality and anchor-policy ergodicity, and conditional on certified oracles for stopping and computing near-minimizing alternative kernels, together with exact occupancy updates, the algorithm is δ-correct and satisfies

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}_p[\tau_\delta]}{\log(1/\delta)} \le T^\star_{2\mu}(p),$$

with vanishing oracle tolerances. Here $T^\star_{2\mu}(p)$ restricts the characteristic-time optimization to occupancies with every coordinate at least 2μ; the bound holds for every fixed floor $\mu > 0$ for which this set is nonempty. It converges to $T^\star(p)$ as $\mu \downarrow 0$, recovering the lower-bound rate in the two-stage vanishing-floor limit. We also compare joint testing with separately optimized experiments and establish the existence of a single floor schedule, independent of the unknown kernel, attaining the exact first-order rate at every admissible instance. We extend the guarantee to certified inexact occupancy updates and implement this variant on a two-state instance, using fewer transitions but more computation than uniform exploration.

1 Introduction

An exploration trajectory can provide evidence about several policies at once. The challenge is to collect that evidence where it is most useful: a state-action pair that resolves one policy's value may contribute little to another. We study *joint policy testing* in discounted tabular Markov decision processes (MDPs). Given a fixed finite family $\Pi_{\mathrm{test}} = \{\pi^{(1)}, \dots, \pi^{(J)}\}$, a known reward, and an unknown transition kernel p, the learner must determine the sign of every value $V_{p,j}(\rho)$ from one adaptive trajectory, with probability at least $1 - \delta$.[1] The target policies are fixed, but the learner may choose a different behavior policy to gather information about them.

Joint testing couples an inference problem with an allocation problem. All policywise decisions use the same transition data, so the exploration strategy must support the hardest remaining sign test. Following each target policy separately can be inefficient: on the two-point reset-chain model class in

[1] Nonzero thresholds can be handled by reward shifting. Policy-specific thresholds lead to policy-specific shifted rewards; we state the theory for a common reward and zero thresholds.

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

Appendix B, on-policy testing has an exponential sampling cost in the chain length L, whereas an active test uses $O_\gamma(L\log(1/\delta))$ transitions. Even with active exploration, the learner cannot prescribe state-action frequencies directly. It chooses actions along a controlled trajectory, and the feasible stationary occupancies depend on the unknown kernel.

Our main contribution is an analysis of how to learn a shared sampling allocation and realize it along one trajectory. The feasible occupancy sets change as the kernel estimate is updated. We control the allocation updates and the resulting sampling frequencies, and prove validity of the stopping tests. The stopping solver is inherited from the fixed-policy framework of Ariu et al. (2026); the new work is to learn and realize the joint allocation while preserving validity under adaptive sampling.

The joint testing lower bound. For each stationary allocation, consider the infimum of the weighted KL divergence over alternatives that reverse at least one policy sign. We maximize this quantity over allocations and denote its inverse by $T^\star(p)$. Theorem 1 shows that every δ-probably correct strategy on a testing class containing all full-row-support instances satisfies, at each instance in that class,

$$\liminf_{\delta\downarrow 0} \frac{\mathbb{E}_p[\tau_\delta]}{\log(1/\delta)} \geq T^\star(p).$$

The characteristic time quantifies the benefit of sharing samples. Proposition 1 compares it with the individual characteristic times: the joint time lies between their maximum and their sum. A common optimal allocation attains the maximum, giving a factor-J improvement over separate experiments when the individual times are equal. Parallel certificates on shared data therefore have a different sampling cost from J independent experiments.

Learning and realizing the allocation. We introduce PT-ACE(μ), short for *Policy Testing via AdaGrad and Cutting-plane Exploration*. Its exchange block learns an allocation using successive cuts over occupancy sets computed from the smoothed empirical kernels. Its navigation block converts averaged exchange centers into an executable behavior policy. Its certification block stops only when the certified lower bounds for all policies are positive on a common plausible set. The technical difficulty is that the feasible domains move, may initially be empty, and must impose a positive floor μ. We control the cut magnitudes, compare the iterates with occupancies of a fixed policy under the changing empirical kernels, and bound the objective at the resulting empirical sampling frequencies. These steps are connected to the stopping-time bound in Section 3.6.

Guarantee and scope. Under nontriviality and anchor-policy ergodicity, Theorem 2 establishes almost-sure termination and δ-correctness for every allowed fixed positive floor, and, whenever the constrained benchmark is feasible,

$$\limsup_{\delta\downarrow 0} \frac{\mathbb{E}_p[\tau_\delta]}{\log(1/\delta)} \leq T^\star_{2\mu}(p), \qquad T^\star_{2\mu}(p) \longrightarrow T^\star(p) \quad (\mu \downarrow 0).$$

Here $T^\star_{2\mu}(p)$ is the characteristic time restricted to stationary occupancies with every coordinate at least 2μ, defined in (3). This sample-complexity guarantee is conditional on the oracle and exact-update requirements in Section 3.5. The floor feasibility condition in (8) depends on the unknown instance. Exact first-order recovery holds in the two-stage limit, first $\delta \downarrow 0$ and then $\mu \downarrow 0$. Corollary 1 strengthens this to one deterministic schedule $\mu(\delta)$, independent of the unknown kernel and fixed within each run, attaining the exact first-order rate at every admissible instance. Its construction is existential; no explicit decay rate is given.

Numerical illustration. Section 4 implements a certified numerical variant, covered by Proposition 14 in Appendix M, on a two-state instance. Both methods use the same certified stopping test. In 100 runs per method, adaptive allocation reduces mean transitions by 49.1%, at higher runtime. Appendix N describes the certified implementation.

1.1 Related work

Fixed-policy testing. Ariu et al. (2026) study testing one policy under a generative model with a static sampling allocation. Their reformulation as optimization with a non-convex objective and a weighted KL constraint, together with the reversed MDP construction, provides the stopping solver used here. We apply that interface separately to each target policy and prove statistical validity for their shared adaptive trajectory. The additional exchange interface is specified in Definition 2.

Online exploration and navigation. Al Marjani et al. (2021) introduce navigation-constrained allocation and tracking for best-policy identification from one trajectory, using non-homogeneous Markov-chain arguments. We build on this online allocation viewpoint. Our objective is to resolve the signs of fixed policies, and the new exchange analysis bounds the joint objective over changing empirical occupancy sets. The lower-bound and allocation approach also connects to best-arm identification (Garivier and Kaufmann, 2016; Kaufmann et al., 2016), game-based pure exploration (Degenne et al., 2019), and Frank–Wolfe exploration (Wang et al., 2021). Thresholding bandits (Locatelli et al., 2016) likewise resolve several signs, but their samples are not constrained by controlled state dynamics.

Multi-policy and multi-reward objectives. Multi-reward best-policy identification (Russo and Vannella, 2024) and online multi-reward multi-policy evaluation (Russo and Pacchiano, 2025) also design one exploration strategy for several downstream questions. The latter seeks value estimates at a prescribed accuracy. Joint policy testing instead stops when every value's sign is resolved, with difficulty determined by the closest sign-flipped alternatives. Our max–min optimization problem is specific to joint sign testing.

2 Problem setup and lower bound

We consider a discounted tabular MDP with finite state space $\mathcal{S}$, finite action space $\mathcal{A}$, discount factor $\gamma \in (0,1)$, unknown transition kernel $p \in \mathcal{P} := \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Delta(\mathcal{S})$, known reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and initial-state distribution $\rho \in \Delta(\mathcal{S})$. A finite family of stationary target policies $\Pi_{\text{test}} := \{\pi^{(1)}, \dots, \pi^{(J)}\}$ is fixed throughout, and $V_{p,j}(\rho) := V_p^{\pi^{(j)}}(\rho)$.

We study joint zero-threshold policy testing. Define $\mathcal{P}_{\text{test}}^{1:J} := \{p \in \mathcal{P} : V_{p,j}(\rho) \neq 0 \text{ for all } j \in [J]\}$. For $p \in \mathcal{P}_{\text{test}}^{1:J}$, let $\text{Ans}_j(p) = +$ if $V_{p,j}(\rho) > 0$ and $\text{Ans}_j(p) = -$ if $V_{p,j}(\rho) < 0$; write $\text{Ans}_{1:J}(p) := (\text{Ans}_1(p), \dots, \text{Ans}_J(p))$. The learner interacts sequentially: at round t, it observes $s(t)$, chooses $a(t)$, and observes $s(t+1)$. A strategy consists of a sampling rule, a stopping time τ, and a decision rule outputting $\hat{\imath} \in \{+,-\}^J$.

2.1 Nontriviality and correct testing

We impose a minimal non-degeneracy condition ensuring that both answers are attainable for every policy in the family. For each $j \in [J]$, write $r^{(j)}(s) := \sum_{a\in\mathcal{A}} \pi^{(j)}(a \mid s) r(s,a)$ and $r^{(j)}(\rho) := \sum_{s\in\mathcal{S}} \rho_s r^{(j)}(s)$. Define $L_j := r^{(j)}(\rho) + \frac{\gamma}{1-\gamma} \min_s r^{(j)}(s)$ and $U_j := r^{(j)}(\rho) + \frac{\gamma}{1-\gamma} \max_s r^{(j)}(s)$.

Assumption 1 (Nontriviality for joint policy testing)**.** *The initial distribution has full support: $\rho_s > 0$ for all $s \in \mathcal{S}$. Moreover, for every $j \in [J]$, $L_j < 0 < U_j$.*

Definition 1 (δ-probably correct joint testing)**.** *For a model class $\mathcal{M} \subseteq \mathcal{P}_{\text{test}}^{1:J}$ and $\delta \in (0,1)$, a strategy is δ-probably correct on $\mathcal{M}$, or δ-PC on $\mathcal{M}$, if for every $p \in \mathcal{M}$,*

$$\mathbb{P}_p(\tau < \infty) = 1, \qquad \mathbb{P}_p(\hat{\imath} \neq \text{Ans}_{1:J}(p)) \leq \delta.$$

For $p \in \mathcal{P}_{\text{test}}^{1:J}$, define $\text{Alt}_j(p) := \{q \in \mathcal{P} : V_{q,j}(\rho) V_{p,j}(\rho) < 0\}$ and $\text{Alt}_{1:J}(p) := \bigcup_{j=1}^J \text{Alt}_j(p)$. Thus $q \in \text{Alt}_{1:J}(p)$ iff at least one target-policy value has the strictly opposite sign under q. Other coordinates may be zero. Appendix C shows that restricting these alternatives to full-row-support kernels in $\mathcal{P}_{\text{test}}^{1:J}$ leaves the information infimum unchanged.

2.2 Occupancy measures and joint characteristic time

An allocation is a probability vector on $\mathcal{S} \times \mathcal{A}$. For a kernel p, let Σ_p be the set of stationary state-action occupancies:

$$\Sigma_p := \left\{ \omega \in \mathbb{R}_+^{\mathcal{S}\times\mathcal{A}} : \sum_{s,a} \omega_{sa} = 1,\ \sum_{b\in\mathcal{A}} \omega_{yb} = \sum_{s,a} \omega_{sa} p(y \mid s,a)\ \forall y \in \mathcal{S} \right\}. \tag{1}$$

Thus ω_{sa} represents the long-run fraction of samples allocated to (s,a), and the equalities in (1) impose stationarity: the total mass assigned to state y equals the total incoming mass generated by p.

For any allocation $\omega \in \Delta(\mathcal{S} \times \mathcal{A})$ and $q \in \mathcal{P}$, write

$$\Phi(p; \omega, q) := \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q), \qquad \mathrm{KL}_{sa}(p, q) := \mathrm{KL}(p(\cdot \mid s, a) \,\|\, q(\cdot \mid s, a)) .$$

We use the convention $0 \cdot (+\infty) = 0$, so rows with zero weight contribute no information. With sampling proportions ω, this is the expected log-likelihood evidence against q per transition under p. For policy j, define

$$F_p^{(j)}(\omega) := \inf_{q \in \mathrm{Alt}_j(p)} \Phi(p; \omega, q),$$

the infimum of the weighted KL divergence over alternatives that reverse the j-th policy sign. A joint decision is incorrect if any policy sign is incorrect, so we minimize over alternatives that reverse at least one sign:

$$F_p^{1:J}(\omega) := \inf_{q \in \mathrm{Alt}_{1:J}(p)} \Phi(p; \omega, q) = \min_{j \in [J]} F_p^{(j)}(\omega).$$

The characteristic time is

$$\frac{1}{T^\star(p)} := \max_{\omega \in \Sigma_p} F_p^{1:J}(\omega). \tag{2}$$

Thus, the allocation maximizes the minimum weighted KL divergence to alternatives that change the joint answer. For a positive floor μ, define the constrained characteristic time

$$T^\star_{2\mu}(p) := \left[\max_{\substack{w \in \Sigma_p \\ w_{sa} \geq 2\mu \ \forall (s,a)}} F_p^{1:J}(w) \right]^{-1}, \tag{3}$$

whenever the constraint set is nonempty. Restricting the allocation gives $T^\star_{2\mu}(p) \geq T^\star(p)$. This is the benchmark used to analyze PT-ACE(μ).

Joint versus separate testing. Define $1/T_j^\star(p) := \max_{\omega \in \Sigma_p} F_p^{(j)}(\omega)$, using the same online occupancy constraint for each individual test.

Proposition 1 (Comparison of characteristic times)**.** *If $0 < T_j^\star(p) < \infty$ for every $j \in [J]$, then*

$$\max_{j \in [J]} T_j^\star(p) \leq T^\star(p) \leq \sum_{j=1}^{J} T_j^\star(p).$$

If the functions $F_p^{(j)}$ have a common maximizer in Σ_p, then $T^\star(p) = \max_j T_j^\star(p)$.

The proof in Appendix C.4 mixes individual optimal occupancies with weights proportional to $T_j^\star(p)$. For fixed J, if separate tests attain their individual constants at confidence δ/J, their total sample count has leading constant $\sum_j T_j^\star(p)$. With a common optimizer and equal individual characteristic times, $T^\star(p)$ is smaller than this sum by a factor J. Policywise certificates evaluated on one shared trajectory instead face the joint allocation problem (2); they are not separate experiments.

2.3 Asymptotic lower bound

Theorem 1 (Asymptotic lower bound for joint testing)**.** *Under Assumption 1, let $\mathcal{M} \subseteq \mathcal{P}_{\mathrm{test}}^{1:J}$ contain every full-row-support kernel in $\mathcal{P}_{\mathrm{test}}^{1:J}$. For every $p \in \mathcal{M}$ and every family of δ-PC strategies on $\mathcal{M}$,*

$$\liminf_{\delta \to 0} \frac{\mathbb{E}_p[\tau]}{\log(1/\delta)} \geq T^\star(p).$$

The proof in Appendix C uses change of measure and expected flow balance; it requires no irreducibility or aperiodicity of p. Theorem 1 identifies the two tasks that any asymptotically optimal algorithm must solve: it must learn an allocation maximizing $F_p^{1:J}$ over Σ_p, as in (2), and it must determine from data when the accumulated evidence rules out every sign-flipped alternative in the family.

3 PT-ACE(μ): certify, navigate, and exchange

The lower bound asks for an allocation maximizing $F_p^{1:J}$ over Σ_p. An online learner cannot directly choose this allocation: it knows neither p nor Σ_p, and it controls actions along a single trajectory. PT-ACE(μ) couples allocation learning with data collection. Each round first asks whether the current data certify every target-policy sign; if not, it collects one transition under a policy chosen before that transition and updates the allocation used for future action selection. Algorithm 1 presents these three blocks in their execution order: *Certify, Navigate, Exchange*.

For the algorithm's termination and navigation guarantees, we now impose an additional condition on the true kernel.

Assumption 2 (Anchor-policy ergodicity)**.** *Let* $\mathrm{Unif}_{\mathcal{A}}$ *denote the uniform policy on* $\mathcal{A}$*. The state-action Markov chain with transition kernel*

$$P_p^{\mathrm{Unif}_{\mathcal{A}}}\big((s,a),(s',a')\big) := p(s' \mid s,a)\mathrm{Unif}_{\mathcal{A}}(a' \mid s')$$

is irreducible and aperiodic.

The upper-bound guarantees use the class

$$\mathcal{P}_{\mathrm{adm}}^{1:J} := \{p \in \mathcal{P}_{\mathrm{test}}^{1:J} : p \text{ satisfies Assumption 2}\}.$$

It contains all full-row-support testing kernels, so Theorem 1 applies with $\mathcal{M} = \mathcal{P}_{\mathrm{adm}}^{1:J}$. Empirical kernels and deterministic oracle problems still range over $\mathcal{P}$.

3.1 State and allocation targets

After t transitions, let $N_{sa}(t)$ be the visit counts, $\hat{w}_{sa}(t) = N_{sa}(t)/(t \vee 1)$ the empirical allocation, and $\hat{p}_t$ the raw empirical kernel (uniform on unobserved rows). The stopping tests use these raw empirical quantities. For allocation learning, we use a full-support surrogate and a positive occupancy floor:

$$\bar{p}_t = (1-\theta_t)\hat{p}_t + \theta_t \mathrm{Unif}_{\mathcal{S}}, \qquad \theta_t = (t \vee 1)^{-1/4}, \qquad K_\mu = \{x \in \Delta(\mathcal{S}\times\mathcal{A}) : x_{sa} \geq \mu \,\forall (s,a)\}.$$

The exchange domain is $\widetilde{\Sigma}_{\bar{p}_t}^{\mu} := \Sigma_{\bar{p}_t} \cap K_\mu$. The learner maintains an exchange center $x_t \in K_\mu$ and a weighted average $\bar{x}_t \in K_\mu$ of corrected or retained centers. The center is used to obtain the next cut; the average determines navigation. In particular, neither object is the observed allocation $\hat{w}(t)$. The averaging weight A_t and the squared-cut accumulator G_t complete the exchange state $\mathsf{E}_t := (x_t, G_t, \bar{x}_t, A_t)$.

The distinction between these allocations matters because the flow equations change with the reference kernel. Centers from different rounds need not satisfy the same flow equations, so their average need not lie in the current stationary occupancy domain. The algorithm uses $\bar{x}_t$ to define action probabilities; the navigation analysis then relates the resulting empirical frequencies to the objective value achieved by the allocation updates.

3.2 Certify: test all signs against one plausible set

The data define a weighted-KL plausible set

$$\mathcal{Q}_t := \{q \in \mathcal{P} : \Phi(\hat{p}_t; \hat{w}(t), q) \leq \sigma_t\}, \qquad \sigma_t := \frac{\beta(t,\delta)}{t \vee 1},$$

with the time-uniform boundary

$$\beta(t,\delta) := \log(1/\delta) + (|\mathcal{S}|-1)\sum_{s,a}\log\left(e\left[1 + \frac{N_{sa}(t)}{|\mathcal{S}|-1}\right]\right). \tag{4}$$

For a target policy j, the optimal value of the non-convex objective (NO) problem

$$u_{\mathrm{NO}}^{(j)}(\sigma; w, p') := \inf_{q:\Phi(p';w,q)\leq\sigma} V_{p',j}(\rho)V_{q,j}(\rho)$$

is positive precisely when every plausible kernel has the same nonzero j-th sign as p'. The certified fixed-policy solver of Ariu et al. (2026) returns

$$L_{t,j}^{\mathrm{stop}} \le u_{\mathrm{NO}}^{(j)}(\sigma_t; \hat{w}(t), \hat{p}_t) \le U_{t,j}^{\mathrm{stop}}, \qquad U_{t,j}^{\mathrm{stop}} - L_{t,j}^{\mathrm{stop}} \le \xi_t^{\mathrm{stop}}.$$

The `Certify` subroutine returns true exactly when all certified lower bounds are positive. It is activated only after every state-action pair has been observed and every empirical target value is nonzero; otherwise it returns false. A bracket containing zero is inconclusive for that round, so the algorithm collects another transition rather than waiting for a strict sign at the same data.

With probability at least $1 - \delta$, the true kernel belongs to $\mathcal{Q}_t$ at every time. Thus simultaneous positivity certifies the whole sign vector on one common event; no division of δ across the J policies is needed. Appendix E gives the deterministic bracket guarantees, and Appendix K proves its validity along the adaptive trajectory.

3.3 Navigate: turn the averaged allocation into actions

Normalize the averaged allocation within each state:

$$\chi_t(a \mid s) := \frac{\bar{x}_t(s,a)}{\sum_b \bar{x}_t(s,b)}.$$

Because $\bar{x}_t \in K_\mu$, every action retains probability at least μ. This persistent exploration supports both learning the kernel and tracking the allocation. The notation distinguishes the navigation policy χ_t from the fixed target policies $\pi^{(j)}$.

We use a one-step delayed convention. The policy χ_t is fixed before $Z_t = (s_t, a_t)$ moves to $Z_{t+1} = (s_{t+1}, a_{t+1})$: after observing s_{t+1}, sample $a_{t+1} \sim \chi_t(\cdot \mid s_{t+1})$. Only then update the empirical kernel and the exchange state. The resulting χ_{t+1} governs the following state-action transition. This timing is the convention used in the navigation proof.

3.4 Exchange: update the sampling allocation

The exchange step alternates between finding an alternative kernel that is hard to distinguish at the current allocation and updating the allocation using the resulting cut. It seeks to maximize the joint objective on $D_t := \widetilde{\Sigma}_{\bar{p}_t}^{\mu}$. To include empirical models with zero target values, use the closed alternatives $\mathrm{Alt}_j^{\le}(p') := \{q \in \mathcal{P} : V_{p',j}(\rho) V_{q,j}(\rho) \le 0\}$ and write

$$F_t(x) := F_{\bar{p}_t}^{1:J}(x) = \min_{j \in [J]} \inf_{q \in \mathrm{Alt}_j^{\le}(\bar{p}_t)} \Phi(\bar{p}_t; x, q).$$

For nonzero reference values and positive allocations, this equals the objective defined using strict alternatives in Section 2.2. As a pointwise infimum of linear functions of x, F_t is concave. Exchange forms successive cuts and applies KL-prox updates. Appendix H bounds the objective of the averaged allocations against a transported comparator.

A cut from an alternative kernel. Suppose D_t and D_{t+1} are nonempty. If the stored center is outside D_t, first reset it to any point in D_t; below, x_t denotes this feasible center. The routine `CutOracle` returns an index j_t and $q_t \in \mathrm{Alt}_{j_t}^{\le}(\bar{p}_t)$ such that

$$g_t(s,a) := \mathrm{KL}_{sa}(\bar{p}_t, q_t), \qquad \langle g_t, x_t \rangle = \Phi(\bar{p}_t; x_t, q_t) \le F_t(x_t) + \xi_t^{\mathrm{cut}}. \tag{5}$$

Thus q_t is a nearly least costly alternative for the joint test, and $g_t(s,a)$ measures the information supplied by sampling row (s,a) against it. Since $F_t(x) \le \Phi(\bar{p}_t; x, q_t) = \langle g_t, x \rangle$,

$$F_t(x) - F_t(x_t) \le \langle g_t, x - x_t \rangle + \xi_t^{\mathrm{cut}} \qquad (x \in D_t).$$

This approximate supergradient inequality shows how the returned kernel guides the allocation update. If a reference value is zero, $F_t \equiv 0$ and $(j_t, q_t, g_t) = (j_t, \bar{p}_t, 0)$ is valid for that policy.

Algorithm 1 PT-ACE(μ): one trajectory, three coupled blocks

Require: confidence δ, floor μ, target policies $\{\pi^{(j)}\}_{j=1}^{J}$, $x_0 \in K_\mu$, tolerances $(\xi_t^{\mathrm{cut}}, \xi_t^{\mathrm{stop}})_{t\geq 0}$

Initialize $N(0) = 0$, $\hat{w}(0) = 0$, $\hat{p}_0 = \bar{p}_0 = \mathrm{Unif}_{\mathcal{S}}$
Initialize $\mathsf{E}_0 = (x_0, G_0, x_0, 0)$, $G_0 > 0$, and normalize x_0 to obtain χ_0
Sample $s_0 \sim \rho$ and $a_0 \sim \chi_0(\cdot \mid s_0)$
for $t = 0, 1, 2, \ldots$ **do**
Certify
  $\sigma_t \leftarrow \beta(t, \delta)/(t \vee 1)$, with β from (4)
  if $\texttt{Certify}(\hat{p}_t, \hat{w}(t), \sigma_t, \xi_t^{\mathrm{stop}})$ **then**
    return $\hat{\imath}_t = \mathrm{Ans}_{1:J}(\hat{p}_t)$
  end if
Navigate
  Observe $s_{t+1} \sim p(\cdot \mid s_t, a_t)$, then sample $a_{t+1} \sim \chi_t(\cdot \mid s_{t+1})$
  Update $N(t+1), \hat{w}(t+1), \hat{p}_{t+1}$ using (s_t, a_t, s_{t+1})
  $\bar{p}_{t+1} \leftarrow (1 - \theta_{t+1})\hat{p}_{t+1} + \theta_{t+1}\mathrm{Unif}_{\mathcal{S}}$
Exchange
  $\mathsf{E}_{t+1} \leftarrow \texttt{ExchangeUpdate}(\bar{p}_t, \bar{p}_{t+1}, \mathsf{E}_t, \mu, \xi_t^{\mathrm{cut}})$
  $\chi_{t+1}(a \mid s) \leftarrow \bar{x}_{t+1}(s, a)/\sum_b \bar{x}_{t+1}(s, b)$
end for

Ascent and correction on moving domains. Adapt the step to the accumulated cut magnitudes: $G_{t+1} = G_t + \|g_t\|_\infty^2$ and $\eta_t = G_{t+1}^{-1/2}$, with $G_0 > 0$. With $\mathrm{KL}(x\|y) := \sum_{s,a} x_{sa} \log(x_{sa}/y_{sa})$, compute

$$\begin{aligned} y_{t+1} &\in \arg\max_{x \in D_t} \left\{ \langle g_t, x \rangle - \eta_t^{-1} \mathrm{KL}(x\|x_t) \right\}, \\ x_{t+1} &\in \arg\min_{x \in D_{t+1}} \mathrm{KL}(x\|y_{t+1}). \end{aligned} \tag{6}$$

The first step favors rows that distinguish q_t, while the KL penalty limits movement from x_t. The second restores feasibility for the updated kernel: the cut and ascent use $\bar{p}_t$, but the corrected center satisfies the flow constraints under $\bar{p}_{t+1}$. If either domain is empty, retain $x_{t+1} = x_t$, set $g_t = 0$, and use the same stepsize rule without an oracle call.

In both branches, average the corrected or retained center:

$$A_{t+1} = A_t + \eta_t, \qquad \bar{x}_{t+1} = \frac{A_t \bar{x}_t + \eta_t x_{t+1}}{A_{t+1}}, \qquad A_0 = 0. \tag{7}$$

This average, rather than the predictor y_{t+1}, determines the next navigation policy. Appendix L gives the full `ExchangeUpdate` subroutine; its oracle and update requirements are collected in Section 3.5.

3.5 Main guarantee

Assumption 2 is navigation-side: together with the persistent action floor induced by K_μ, it gives a common Doeblin minorization for all navigation policies generated by PT-ACE(μ). It is not used by the policywise stopping certificates. Weakening it would require a different navigation argument, such as explicit forced exploration for merely communicating or weakly communicating MDPs.

For the rate bound, require a feasible comparison occupancy with floor 2μ:

$$\Sigma_p \cap K_{2\mu} \neq \varnothing. \tag{8}$$

Assumption 2 ensures this condition for all sufficiently small $\mu > 0$, since the uniform-action policy has a positive stationary occupancy. The algorithm is defined for any $0 < \mu \leq 1/(|\mathcal{S}||\mathcal{A}|)$; only its rate bound needs (8). The definition of $T^\star(p)$ imposes no occupancy floor.

Oracle and update requirements. We require the cut certificate (5) and stopping brackets of width at most ξ_t^{stop}, with deterministic tolerances satisfying

$$\sup_{t\geq 0} \xi_t^{\mathrm{cut}} < \infty, \qquad \xi_t^{\mathrm{cut}} \to 0, \qquad \xi_t^{\mathrm{stop}} \to 0. \tag{9}$$

Every oracle query must terminate with the stated certificate, and the two occupancy KL-prox updates in Algorithm 4 are solved exactly. Appendix E obtains the stopping brackets from the scalar solver of Ariu et al. (2026). For nonzero reference values, the exchange step additionally requires the feasible alternative kernel and objective certificate of Definition 2; a scalar bracket alone does not supply the kernel. Appendix F proves the reduction from that requirement to (5). The theorem is conditional on these interfaces and exact updates; it does not assert an end-to-end polynomial runtime bound.

Theorem 2 (Correctness and asymptotic upper bound for PT-ACE(μ))**.** *Suppose Assumption 1 holds. Fix $0 < \mu \le 1/(|\mathcal{S}||\mathcal{A}|)$, and run PT-ACE($\mu$) using the oracles specified above and exact occupancy updates. Use positive deterministic tolerance schedules $(\xi_t^{\mathrm{cut}})_{t\ge 0}$ and $(\xi_t^{\mathrm{stop}})_{t\ge 0}$ satisfying* (9)*, held fixed as δ varies. Then PT-ACE(μ) is δ-PC on $\mathcal{P}_{\mathrm{adm}}^{1:J}$ for every $\delta \in (0,1)$. For each $p \in \mathcal{P}_{\mathrm{adm}}^{1:J}$ such that $\Sigma_p \cap K_{2\mu} \neq \varnothing$, it also satisfies*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_p[\tau_\delta]}{\log(1/\delta)} \le T^\star_{2\mu}(p).$$

Since $T^\star_{2\mu}(p) \to T^\star(p)$ as $\mu \downarrow 0$, $\{\mathrm{PT\text{-}ACE}(\mu)\}_{\mu>0}$ recovers the lower-bound characteristic time in the two-stage limit: first $\delta \downarrow 0$ for fixed μ satisfying (8)*, and then $\mu \downarrow 0$.*

If an unconstrained optimal occupancy has every coordinate at least 2μ, then $T^\star_{2\mu}(p) = T^\star(p)$: the fixed floor already attains the first-order lower bound. Otherwise the constrained benchmark may be larger. Smaller floors reduce this benchmark, while the mixing, burn-in, and numerical constants in the proof can deteriorate. Both the feasibility condition and these constants depend on the unknown instance; the theorem gives no uniform-in-μ rate.

Corollary 1 (Asymptotic optimality)**.** *Under Assumption 1, use the implementation required by Theorem 2, with tolerance schedules fixed independently of p and δ. There is a deterministic schedule $0 < \mu(\delta) \le 1/(|\mathcal{S}||\mathcal{A}|)$, independent of p, with $\mu(\delta) \to 0$ as $\delta \downarrow 0$, such that*

$$\lim_{\delta\downarrow 0} \frac{\mathbb{E}_p[\tau_{\delta,\mu(\delta)}]}{\log(1/\delta)} = T^\star(p) \qquad \text{for every } p \in \mathcal{P}_{\mathrm{adm}}^{1:J}.$$

Here $\tau_{\delta,\mu}$ is the stopping time of PT-ACE(μ) at confidence δ, with the floor fixed throughout each run. The class $\mathcal{P}_{\mathrm{adm}}^{1:J}$ incorporates Assumption 2. The schedule is an existence result, with pointwise convergence in p; an arbitrary decay rate is not justified.

Certified numerical implementation. Proposition 14 in Appendix M preserves these conclusions with exactly feasible occupancy updates having summable certified first-order errors. It also allows certified arithmetic for cuts, averaging, and stopping.

3.6 Proof sketch

Proof sketch of Theorem 2. The argument proceeds in four steps.

1. **Correctness and termination.** On the common time-uniform confidence event, the true kernel belongs to every plausible set. The certified stopping rule therefore returns all policy signs correctly. For every fixed positive floor, persistent exploration and empirical consistency also imply almost-sure termination. The confidence argument and Lemma 23 in Appendix K.1 establish these claims.
2. **Learning an informative allocation.** Fix μ with $\Sigma_p \cap K_{2\mu} \neq \varnothing$. Choose an optimal constrained occupancy $w^{2\mu}$, normalize it to obtain a fixed policy, and transport that policy to the stabilized empirical kernels. Lemma 13 in Appendix H shows that the transported occupancies remain feasible and have controlled cumulative movement: the slack between 2μ and μ absorbs the occupancy perturbation caused by kernel estimation.

   Assumption 1 ensures that each target policy can have either a strictly positive or a strictly negative value. Continuity allows us to choose fixed, full-support kernels realizing these two signs. Their rowwise KL costs are uniformly bounded over all reference kernels, giving $F_t(x_t) \le C^\circ < \infty$. Consequently,

$$\mu\|g_t\|_\infty \le \langle x_t, g_t\rangle \le C^\circ + \sup_u \xi_u^{\mathrm{cut}}.$$

This is the bounded-cut argument in Lemma 16 in Appendix H. Summing the one-step inequality of Lemma 17 in Appendix H along the transported comparators makes the KL terms telescope. After normalization, comparator movement and optimization errors vanish. Continuity and averaging then yield the objective bound in Proposition 8 in Appendix H.

3. **Transferring information to the trajectory.** The action floor supplies common mixing, and averaging makes the navigation policies vary slowly. Proposition 10 in Appendix I controls empirical frequencies around average stationary occupancies. Lemma 20 in Appendix I connects these occupancies to the exchange averages through their flow error. Combining these results gives Proposition 11 in Appendix I: the empirical allocation approaches the same information benchmark, without an additional first-order loss.

4. **Bounding the expected stopping time.** Kernel concentration and continuity transfer this benchmark to the empirical stopping objective. For every $\lambda > T^{\star}_{2\mu}(p)$, the resulting information bound eventually exceeds the normalized stopping threshold after $\lambda \log(1/\delta)$ transitions. Lemma 24 in Appendix K shows that the certified stopping rule then terminates. The tail bound in Lemma 25 in Appendix K controls the exceptional trajectories; summing it proves the expected-time guarantee.

Proof sketch of Corollary 1. Lemma 11 in Appendix H gives $T^{\star}_{2\mu}(p) \to T^{\star}(p)$ as $\mu \downarrow 0$. To select a single floor schedule independent of p, Lemma 26 in Appendix K.4 first makes the fixed-floor upper bound uniform on nested compact classes. Appendix K.4 then diagonalizes over these classes, floors, and confidence thresholds. Theorem 1 supplies the matching lower bound. This establishes the existence of a common schedule; it does not justify an arbitrary rate of floor decay.

3.7 Computational considerations

Write $n = |\mathcal{S}|$, $m = |\mathcal{A}|$, and $d = nm$. A round with active stopping and exchange uses J scalar stopping solves at width ξ_t^{stop} and, when reference values are nonzero, J certified alternative-kernel computations at accuracy $\xi_t^{\mathrm{cut}}/4$. Each optimization ranges over a transition kernel with $n^2 m$ coordinates. With current-domain feasibility cached, exchange additionally requires one feasibility LP and two occupancy KL-prox solves in d variables. Forming the cut costs $O(n^2 m)$ arithmetic operations. These counts describe the interfaces in Section 3.5; they do not establish an end-to-end polynomial runtime bound.

For the raw empirical model, each transition changes only one row. Sherman–Morrison updates therefore maintain all J target-policy values in $O(Jn^2)$ operations per transition after $O(Jn^3)$ initialization; Appendix L.4 derives the update. These values determine the empirical signs used by stopping. This maintenance cost excludes certified optimization and policy evaluation at the stabilized model, whose changing smoothing weight can modify every row.

4 Numerical illustration

We compare adaptive allocation with uniform-action exploration on a two-state, two-action MDP. Let $\mathcal{S} = \{0, 1\}$, $\mathcal{A} = \{a_0, a_1\}$, $\gamma = 0.9$, and $\rho = (1/2, 1/2)$. The reward is $r(0, a) = -1/2$, $r(1, a) = 1/2$ for either action, and

$$p(1 \mid s, a_0) = 0.58, \qquad p(1 \mid s, a_1) = 0.12 \qquad (s \in \mathcal{S}).$$

The target policies always choose a_0 and a_1, with values $V_{p,1}(\rho) = 0.72$ and $V_{p,2}(\rho) = -3.42$, respectively (see (22)). The first policy is closer to its testing threshold, so this instance tests whether exploration shifts toward a_0.

Protocol. We use $\delta = 0.1$, $\mu = 0.02$, and 100 independent trajectories per method. PT-ACE updates after every transition, using closed-form alternatives, exact rational feasibility, and certified projection errors. The baseline chooses actions uniformly; both apply the same interval-certified stopping test; see Appendix N for implementation details. Code and verification records are provided in the supplement.

Table 1: Two-state illustration: 100 runs per method at $\delta = 0.1$. Stopping times count transitions; parentheses give sample standard deviations divided by $\sqrt{100}$. Errors count incorrect joint sign decisions. Action fractions and runtimes are averaged over runs; timings include certification.

Method	Errors	Mean stop (SE)	a_0 fraction	Time/run
PT-ACE (certified)	0/100	3035 (87)	0.805	5.69 s
Uniform action	0/100	5966 (169)	0.500	3.61 s

Results and scope. Adaptive allocation increases the mean a_0 fraction from 0.500 to 0.805 and reduces mean stopping time by 49.1%, at higher runtime. Figure 2 shows the policywise evidence. Correctness follows from Proposition 14 in Appendix M, while this empirical comparison covers only one instance and does not establish a general sampling advantage.

5 Conclusion

Joint policy testing shares transition data across sign decisions, with first-order complexity characterized by $T^\star(p)$. PT-ACE(μ) learns allocations over changing empirical occupancy sets and realizes them along one trajectory. Under Theorem 2's conditions, its asymptotic upper bound is $T^\star_{2\mu}(p)$, converging to $T^\star(p)$ as the floor vanishes. The numerical illustration shows the potential sampling benefit and computational cost.

Floor feasibility depends on the unknown instance; the common schedule in Corollary 1 has existential confidence thresholds. The guarantee is asymptotic and requires anchor-policy ergodicity and certified optimization; the numerical extension requires exactly feasible updates with controlled errors. The experiment covers one small instance, and no end-to-end polynomial runtime bound is established.

These limitations motivate further work on explicit confidence-dependent or within-run floor schedules, weaker navigation assumptions, finite-δ guarantees, and more efficient certified optimization. The allocation, navigation, and stopping framework may also extend to other pure-exploration problems in MDPs.

Policy-value certificates may support evidence-based decisions, but certify only the specified reward, policies, and model class. They do not ensure social desirability, fairness, or robustness to unmodeled shifts. High-stakes applications require domain validation, safety constraints, and human oversight.

Acknowledgements

We thank the anonymous reviewers for their constructive reviews. Po-An Wang is supported by NSTC Grant Number 114-2118-M-007-002-MY2. Kaito Ariu is supported by JST BOOST Program Japan Grant Number JPMJBY25A0 and JSPS KAKENHI Grant Number JP25K21291. The authors declare no additional competing interests.

References

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

Ariu, K., Wang, P.-A., Proutiere, A., and Abe, K. (2026). Policy testing in Markov decision processes. In Khan, E., Li, Y., Solin, A., and Ramdas, A., editors, *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 568–576. PMLR.

Degenne, R., Koolen, W. M., and Ménard, P. (2019). Non-asymptotic pure exploration by solving games. In *Advances in Neural Information Processing Systems*, volume 32.

Fort, G., Moulines, E., and Priouret, P. (2011). Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *The Annals of Statistics*, 39(6):3262–3289.

Garivier, A. and Kaufmann, E. (2016). Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 998–1027. PMLR.

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. (2020). Array programming with NumPy. *Nature*, 585(7825):357–362.

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42.

Locatelli, A., Gutzeit, M., and Carpentier, A. (2016). An optimal algorithm for the thresholding bandit problem. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1690–1698. PMLR.

Russo, A. and Pacchiano, A. (2025). Adaptive exploration for multi-reward multi-policy evaluation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 52382–52421. PMLR.

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

Wang, P.-A., Tzeng, R.-C., and Proutiere, A. (2021). Fast pure exploration via Frank–Wolfe. In *Advances in Neural Information Processing Systems*, volume 34, pages 5810–5821.

Contents

1 Introduction **1**

1.1 Related work 2

2 Problem setup and lower bound **3**

2.1 Nontriviality and correct testing 3

2.2 Occupancy measures and joint characteristic time 3

2.3 Asymptotic lower bound 4

3 PT-ACE(μ): certify, navigate, and exchange **5**

3.1 State and allocation targets 5

3.2 Certify: test all signs against one plausible set 5

3.3 Navigate: turn the averaged allocation into actions 6

3.4 Exchange: update the sampling allocation 6

3.5 Main guarantee 7

3.6 Proof sketch 8

3.7 Computational considerations 9

4 Numerical illustration **9**

5 Conclusion **10**

A Notation table **14**

B A toy instance where on-policy testing is exponentially inefficient **14**

C Sample-complexity lower bound for simultaneous testing **16**

C.1 Constraints on expected visit counts 16

C.2 Full-support testing alternatives suffice 17

C.3 Normalization and the characteristic-time bound 17

C.4 Joint and individual characteristic times 18

D Two optimization formulations for fixed-policy testing **18**

E Using the fixed-policy solver for stopping **19**

F Certified allocation cuts **20**

F.1 Certified alternative kernels 20

F.2 From policywise alternatives to a joint allocation cut 20

G Concentration of the empirical kernels **21**

G.1 The action floor and common minorization 21

G.2 Linear visit counts with exponential tails 22

G.3 Empirical-kernel concentration 22

H Exchange analysis for PT-ACE(μ) 23

H.1 Objective and exchange update 23

H.2 A feasible comparator on the changing domains 24

H.3 A common continuity lemma for the KL objectives 25

H.4 The one-step inequality and its cumulative error 26

H.5 Completing the objective bound and controlling flow 27

I Navigation 28

I.1 Slow variation and uniform Poisson bounds 29

I.2 Tracking the changing stationary occupancies 30

I.3 From flow balance to the empirical objective 30

J Continuity of KL minima and positive bounds for stopping 31

J.1 Positive empirical allocations and local continuity 31

J.2 A uniform positive stopping margin 32

K Proof of Theorem 2 32

K.1 Termination and the δ-probably correct guarantee 32

K.2 Deterministic stopping on the good event 33

K.3 Stopping tails and the expectation bound 34

K.4 A floor schedule independent of the unknown kernel 34

L Algorithm subroutines and online computation 36

L.1 Empirical objects and initialization 36

L.2 The certification subroutine 36

L.3 The exchange subroutine 36

L.4 Rank-one maintenance of raw empirical policy values 37

M Extension to certified numerical updates 38

N Experimental Details 39

N.1 Policy values and KL minimization 39

N.2 Certified allocation updates 40

N.3 Sampling and computational resources 41

N.4 Policywise evidence 41

NeurIPS Paper Checklist 41

Appendix roadmap

The appendices are organized as follows. Appendix A collects frequently used notation. Appendix B gives a reset-chain example showing why on-policy testing can be exponentially inefficient. Ap-

pendix C proves the sample-complexity lower bound. Appendix D relates the two optimization formulations for testing each policy. Appendix E explains how the certified policywise stopping oracle is instantiated from the fixed-policy testing solver of Ariu et al. (2026). Appendix F states the additional requirements for certified alternative kernels and derives the exchange cuts. Appendix G proves adaptive kernel concentration. Appendix H analyzes allocation updates over changing empirical occupancy sets. Appendix I relates averaged target occupancies to empirical sampling frequencies. Appendix J establishes continuity of the KL minimization values and a positive lower bound on the stopping objective. Appendix K proves the main upper bound. Appendix L gives the algorithm subroutines and online value updates. Appendix M extends the guarantee to certified numerical occupancy updates. Appendix N gives implementation details and policywise evidence for the numerical experiment.

A Notation table

Symbol	Meaning
$\mathcal{S}, \mathcal{A}$	finite state and action spaces
γ, ρ, r	discount factor, initial distribution, and known reward function
$p, \hat{p}_t, \bar{p}_t$	true kernel, empirical kernel, and stabilized empirical kernel
Π_{test}	finite family of target policies $\{\pi^{(j)}\}_{j=1}^{J}$
$\mathcal{P}_{\mathrm{test}}^{1:J}$	kernels with nonzero values for all target policies
$\mathcal{M}$	model class on which termination and correctness are required
$\mathcal{P}_{\mathrm{adm}}^{1:J}$	anchor-ergodic testing class for the upper-bound guarantees
$V_{p,j}(\rho)$	value of target policy $\pi^{(j)}$ under kernel p
$\mathrm{Ans}_{1:J}(p)$	vector of target-policy signs under p
$\mathrm{Alt}_j(p), \mathrm{Alt}_{1:J}(p)$	policywise and joint sign-flipped alternatives
Σ_p	stationary state-action occupancy set under kernel p
K_μ	allocation set $\{w : w_{sa} \geq \mu \; \forall (s,a)\}$
$\widetilde{\Sigma}_{p'}^{\mu}$	fixed-floor occupancy domain $\Sigma_{p'} \cap K_\mu$
$\Phi(p; \omega, q)$	weighted KL divergence $\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p, q)$
$F_p^{(j)}, F_p^{1:J}$	KL objective for policy j, and its minimum over policies
$T^\star(p)$	characteristic time, $1/T^\star(p) = \max_{\omega \in \Sigma_p} F_p^{1:J}(\omega)$
$T_{2\mu}^\star(p)$	characteristic time with occupancy floor 2μ, (3)
$x_t, y_{t+1}, \bar{x}_t$	exchange center, predictor, and averaged corrected center
χ_t	policy used for the next transition, obtained from $\bar{x}_t$
g_t, q_t, j_t	cut vector, returned alternative kernel, and selected policy index
$\xi_t^{\mathrm{cut}}, \xi_t^{\mathrm{stop}}$	cut and stopping numerical tolerances
$\hat{w}(t), N_{sa}(t)$	empirical allocation and state-action visit counts
$\mathcal{C}_T^{\mathrm{ker}}, \mathcal{C}_T^{\mathrm{nav}}$	kernel and navigation good events
τ_T, ϑ_T	starting indices of the analysis intervals $\lceil T^{1/4} \rceil$, $\lceil T^{3/4} \rceil$
τ_δ	stopping time of PT-ACE(μ)

Table 2: Frequently used notation.

B A toy instance where on-policy testing is exponentially inefficient

This two-point example shows how active navigation efficiently reaches an informative row that the target policy visits exponentially rarely. It is illustrative and is not used in the main theorem; in particular, its initial distribution is the point mass $\rho = \delta_{s_0}$.

The reset chain. Fix $\gamma \in (0,1)$ and $L \geq 1$. The states are $s_0, \ldots, s_L$. For $i < L$, action a advances to s_{i+1} with probability p_i and otherwise resets to s_0; action b advances deterministically. State s_L gives reward one and resets deterministically to s_0. All other rewards are zero. The target policy π always chooses a, with the sole reset action at s_L.

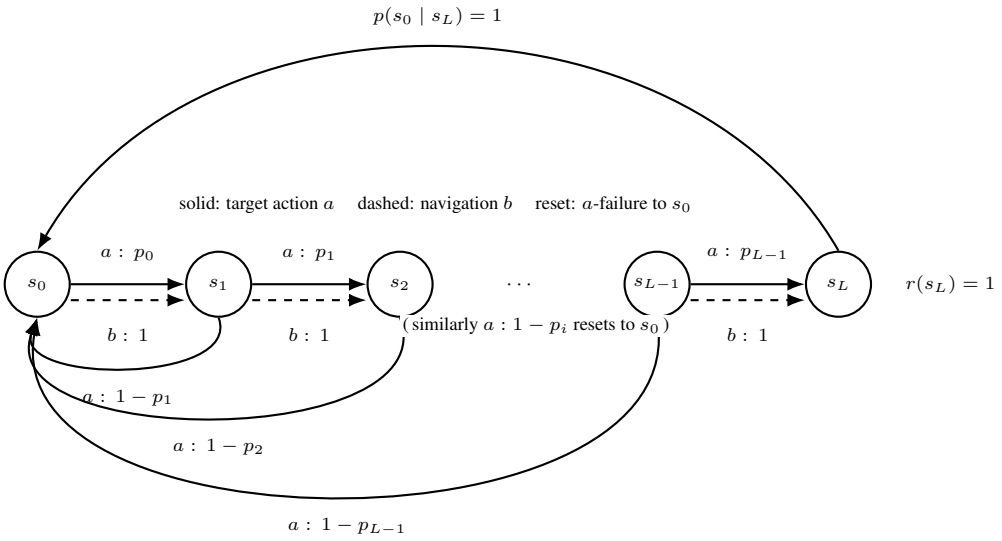


Figure 1: The reset-chain instance. State s_L is a rewarding reset state: it yields reward 1 and then deterministically returns to s_0.

Lemma 1 (Value of the target policy)**.** *Put* $\alpha_i := \prod_{j<i} p_j$*, with* $\alpha_0 = 1$*, and* $d_p := \sum_{i=0}^{L-1} \gamma^{i+1}\alpha_i(1-p_i) + \gamma^{L+1}\alpha_L$*. Then*

$$V_p^\pi(s_0) = \frac{\gamma^L \alpha_L}{1 - d_p}, \qquad \gamma^L \alpha_L \le V_p^\pi(s_0) \le \frac{\gamma^L \alpha_L}{1-\gamma}.$$

Proof. A cycle from s_0 to its next return either fails at level $i < L$, after $i+1$ transitions, or reaches s_L, earns reward at time L, and resets at time $L+1$. Its expected discounted reward is $\gamma^L \alpha_L$, and its expected discount at reset is d_p. The renewal identity $V = \gamma^L \alpha_L + d_p V$ gives the formula. Every cycle has length at least one, so $0 \le d_p \le \gamma$, giving both bounds. □

Testing whether this value exceeds a known $r_\star$ is a zero-threshold test after shifting the reward to $\tilde{r}(s,a) = r(s,a) - (1-\gamma)r_\star$, since $V_{p,\tilde{r}}^\pi(s_0) = V_p^\pi(s_0) - r_\star$.

Lemma 2 (On-policy visits to the informative row)**.** *If* $p_i = 1/2$ *for every* $i < L$*, then any on-policy stopping time* T *satisfies*

$$\mathbb{E}[N_{L-1}(T)] \le 2^{-(L-1)}(\mathbb{E}[T] + 1), \qquad N_{L-1}(T) := \sum_{t<T} \mathbf{1}\{S_t = s_{L-1}, A_t = a\}.$$

Proof. Extend the on-policy trajectory after stopping. Let R_k indicate that cycle k reaches s_{L-1}, and let K_T count cycles started before T. Each R_k has mean $2^{-(L-1)}$, independently of the history at its cycle's start. The event $\{K_T \ge k\}$ is measurable in that history. Each cycle visits the row at most once, and $K_T \le T + 1$, so Tonelli's theorem and conditional expectation give

$$\mathbb{E}[N_{L-1}(T)] \le \sum_{k \ge 1} \mathbb{E}[\mathbf{1}\{K_T \ge k\} R_k] = 2^{-(L-1)} \mathbb{E}[K_T] \le 2^{-(L-1)}(\mathbb{E}[T] + 1).$$

□

Proposition 2 (Exponential versus linear sampling cost)**.** *For every* $\gamma \in (0,1)$ *and* $L \ge 1$*, there are two reset-chain kernels* p^+, p^- *and a threshold* $r_\star$ *with* $V_{p^+}^\pi(s_0) > r_\star > V_{p^-}^\pi(s_0)$*. There are constants* $c_\gamma, C_\gamma > 0$*, independent of* L *and* δ*, such that, for* $0 < \delta \le 1/4$*,*

$$\mathbb{E}_{p^+}[T] \ge c_\gamma 2^L \log(1/\delta) - 1 \quad \textit{for every on-policy test correct on both kernels},$$
$$\mathbb{E}_{p^+}[T] \vee \mathbb{E}_{p^-}[T] \le C_\gamma L \log(1/\delta) \quad \textit{for an active test on the two-point model class}.$$

Here correctness means almost-sure termination and error probability at most δ *under each kernel.*

Proof. Set $p_i^+ = 1/2$ for all $i < L$, and change only the last row under p^-, to $p^-_{L-1} = r_\gamma := (1-\gamma)/8$. The value lemma gives

$$V^\pi_{p^-}(s_0) \le \tfrac{1}{4} V^\pi_{p^+}(s_0), \qquad r_\star := \tfrac{1}{2}\big(V^\pi_{p^+}(s_0) + V^\pi_{p^-}(s_0)\big),$$

so the correct answers differ.

On-policy lower bound. Only (s_{L-1}, a) distinguishes the kernels. Write $D_\gamma := \mathrm{kl}(1/2, r_\gamma)$ for their Bernoulli KL divergence. If $\mathbb{E}_{p^+}[T] = \infty$, the claim is immediate. Otherwise the stopped likelihood-ratio identity, binary change of measure, and the visit lemma give

$$\mathrm{kl}(1-\delta, \delta) \le D_\gamma \mathbb{E}_{p^+}[N_{L-1}(T)] \le D_\gamma 2^{-(L-1)}(\mathbb{E}_{p^+}[T] + 1).$$

This exact inequality holds for every $0 < \delta < 1/2$. For $\delta \le 1/4$, $\mathrm{kl}(1-\delta, \delta) \ge \frac{1}{4}\log(1/\delta)$, giving the stated bound with $c_\gamma = 1/(8D_\gamma)$.

Active upper bound. Navigate from s_0 to s_{L-1} using b, then play a once. Record whether the next state is s_L, and reset to s_0 if needed. Each probe costs at most $L+1$ transitions. The successive probe outcomes are independent Bernoulli variables, with mean either $1/2$ or r_γ. Put $\Delta_\gamma := 1/2 - r_\gamma > 0$ and take

$$n = \left\lceil \frac{2}{\Delta_\gamma^2} \log(1/\delta) \right\rceil$$

probes. Comparing their mean with $(1/2 + r_\gamma)/2$, Hoeffding's inequality bounds either error probability by $\exp(-n\Delta_\gamma^2/2) \le \delta$. The deterministic cost bound $T \le (L+1)n \le C_\gamma L \log(1/\delta)$, for $\delta \le 1/4$, proves the claim. □

The active strategy uses knowledge of the two possible kernels. The separation shows why on-policy sampling can be costly even in this simple experiment; the main results analyze the general problem with an unknown kernel and several target policies.

C Sample-complexity lower bound for simultaneous testing

We prove Theorem 1 by normalizing expected visit counts. Change of measure imposes information constraints, while a telescoping identity makes any limiting allocation stationary. A full-support approximation connects these constraints to the characteristic time, including at kernels with zero transition probabilities.

Assume Assumption 1. Fix $\mathcal{M} \subseteq \mathcal{P}^{1:J}_{\mathrm{test}}$ containing every full-row-support testing kernel, and $p \in \mathcal{M}$. No ergodicity assumption is used. Retain the main-text definitions of $\mathrm{Alt}_j(p)$, $\mathrm{Alt}_{1:J}(p)$, $\Phi(p; \omega, q)$, and Σ_p, with $0 \cdot \infty = 0$. For change of measure, use

$$\mathrm{Alt}^\circ_{1:J}(p) := \{q \in \mathcal{P}^{1:J}_{\mathrm{test}} : \mathrm{Ans}_{1:J}(q) \ne \mathrm{Ans}_{1:J}(p), \quad q(y \mid s, a) > 0 \; \forall (s, a, y)\}.$$

These alternatives belong to $\mathcal{M}$ and have finite rowwise KL costs. They exclude both zero transition probabilities and zero values of other target policies; the approximation below shows that neither exclusion changes the information infimum.

C.1 Constraints on expected visit counts

After t transitions, use $N_{sa}(t) := \sum_{u=0}^{t-1} \mathbf{1}\{S_u = s, A_u = a\}$ and $N_y(t) := \sum_a N_{ya}(t)$.

Lemma 3 (Expected flow balance; Al Marjani et al., 2021)**.** *For any strategy with $\mathbb{E}_p[\tau] < \infty$,*

$$\left| \mathbb{E}_p[N_y(\tau)] - \sum_{s,a} p(y \mid s, a) \mathbb{E}_p[N_{sa}(\tau)] \right| \le 1 \qquad (y \in \mathcal{S}).$$

Proof. Conditional expectation and telescoping give the difference inside the absolute value as $\mathbb{E}_p[\mathbf{1}\{S_0 = y\} - \mathbf{1}\{S_\tau = y\}]$. Finite expected stopping time justifies summing the conditional expectations, and the boundary term lies in $[-1, 1]$. □

Lemma 4 (Information constraint)**.** *For $0 < \delta < 1/2$, a δ-PC strategy on $\mathcal{M}$ with $\mathbb{E}_p[\tau] < \infty$ satisfies, for every $q \in \mathrm{Alt}^\circ_{1:J}(p)$,*

$$\sum_{s,a} \mathbb{E}_p[N_{sa}(\tau)]\mathrm{KL}_{sa}(p,q) \geq k_\delta := \mathrm{kl}(1-\delta,\delta).$$

Proof. For the stopped event $E = \{\hat{\imath} = \mathrm{Ans}_{1:J}(p)\}$, correctness on $\mathcal{M}$ gives $\mathbb{P}_p(E) \geq 1-\delta$ and $\mathbb{P}_q(E) \leq \delta$. The stopped likelihood-ratio identity and data processing give

$$\sum_{s,a} \mathbb{E}_p[N_{sa}(\tau)]\mathrm{KL}_{sa}(p,q) \geq \mathrm{kl}(\mathbb{P}_p(E), \mathbb{P}_q(E)) \geq \mathrm{kl}(1-\delta,\delta).$$

The final inequality uses monotonicity of binary KL on these probability ranges. Full support of q makes every rowwise divergence finite. □

C.2 Full-support testing alternatives suffice

Lemma 5 (Full-support approximation)**.** *For every $\omega \in \Delta(\mathcal{S} \times \mathcal{A})$,*

$$\inf_{q \in \mathrm{Alt}_{1:J}(p)} \Phi(p;\omega,q) = \inf_{q \in \mathrm{Alt}^\circ_{1:J}(p)} \Phi(p;\omega,q).$$

The same equality holds for each policywise alternative set and its full-support testing subset.

Proof. The full-support subset gives the easy inequality. For the reverse, fix any $q \in \mathrm{Alt}_{1:J}(p)$ of finite cost, reversing policy j_0. Choose a full-support testing kernel q° near p: a sufficiently small mixture of p with the rowwise uniform kernel preserves all its nonzero policy values. Set

$$q_\varepsilon := (1-\varepsilon)q + \varepsilon q^\circ, \qquad 0 < \varepsilon \leq 1.$$

This mixture has full support and still reverses j_0 for small ε, by value continuity. For every k,

$$V_{q_\varepsilon,k}(\rho) = \rho^\top (I - \gamma P_{q_\varepsilon}^{\pi^{(k)}})^{-1} r^{(k)}$$

is a rational function of ε, with denominator nonzero on $[0,1]$. Its numerator is a polynomial that is nonzero at $\varepsilon = 1$, so it has only finitely many zeros. Since J is finite, all sufficiently small positive ε avoid every such zero. Thus $q_\varepsilon \in \mathrm{Alt}^\circ_{1:J}(p)$ for those ε.

On each positively weighted row, finite KL cost means that q is positive wherever p is positive. Rowwise KL therefore converges as $q_\varepsilon \to q$; zero-weight rows contribute nothing. Hence

$$\Phi(p;\omega,q_\varepsilon) \longrightarrow \Phi(p;\omega,q).$$

Taking infima proves equality. The argument preserves the chosen reversed policy j_0, so it also proves the policywise statement. □

In particular, the joint and policywise objectives are infima of finite linear functions of ω. They are upper semicontinuous, and therefore attain their maxima on the compact occupancy set Σ_p.

C.3 Normalization and the characteristic-time bound

Proof of Theorem 1. Since $k_\delta / \log(1/\delta) \to 1$, it suffices to prove

$$M := \liminf_{\delta \downarrow 0} \frac{\mathbb{E}_p[\tau]}{k_\delta} \geq T^\star(p).$$

If $M = \infty$, the claim is immediate. Otherwise choose $\delta_m \downarrow 0$ along which the finite normalized expectations converge to M, and put

$$n_{sa}^{(m)} := \frac{\mathbb{E}_p[N_{sa}(\tau_{\delta_m})]}{k_{\delta_m}}, \qquad \sum_{s,a} n_{sa}^{(m)} \longrightarrow M.$$

These nonnegative vectors are bounded. Pass to a convergent subsequence $n^{(m)} \to n$. The flow and information lemmas give

$$\left|\sum_a n_{ya}^{(m)} - \sum_{s,a} p(y \mid s,a) n_{sa}^{(m)}\right| \le k_{\delta_m}^{-1} \longrightarrow 0,$$
$$\sum_{s,a} n_{sa}^{(m)} \mathrm{KL}_{sa}(p,q) \ge 1 \qquad (q \in \mathrm{Alt}_{1:J}^{\circ}(p)).$$

For each fixed full-support q, its rowwise KL costs are finite, so both constraints pass to the limit. Nontriviality and the approximation lemma ensure that this alternative set is nonempty; the limiting information constraint therefore forces $M = \sum_{s,a} n_{sa} > 0$. Normalize $\omega = n/M$. The limiting flow equations imply $\omega \in \Sigma_p$, and

$$\frac{1}{M} \le \inf_{q \in \mathrm{Alt}_{1:J}^{\circ}(p)} \Phi(p;\omega,q) = F_p^{1:J}(\omega) \le \max_{w \in \Sigma_p} F_p^{1:J}(w) = \frac{1}{T^\star(p)}.$$

Thus $M \ge T^\star(p)$. If the maximum is zero, the display rules out any finite M, giving the same conclusion with $T^\star(p) = \infty$. This proves the theorem without any mixing assumption. □

C.4 Joint and individual characteristic times

Proof of Proposition 1. Write $C_j = 1/T_j^\star(p)$ and $C = 1/T^\star(p)$. Since $\min_k F_p^{(k)}(\omega) \le F_p^{(j)}(\omega)$, maximizing over Σ_p gives $C \le \min_j C_j$, hence $T^\star(p) \ge \max_j T_j^\star(p)$.

For the reverse inequality, choose individual maximizers $\omega_j^\star$, put $S_p = \sum_j T_j^\star(p)$, and mix them with weights $\lambda_j = T_j^\star(p)/S_p$. Convexity gives $\bar\omega := \sum_j \lambda_j \omega_j^\star \in \Sigma_p$. Concavity and nonnegativity yield, for every j,

$$F_p^{(j)}(\bar\omega) \ge \sum_k \lambda_k F_p^{(j)}(\omega_k^\star) \ge \lambda_j C_j = 1/S_p.$$

Thus $C \ge 1/S_p$ and $T^\star(p) \le S_p$. If a common maximizer exists, it attains $\min_j C_j$ for the joint objective, giving $T^\star(p) = \max_j T_j^\star(p)$. □

For fixed J, separate tests attaining their individual constants at confidence δ/J have total leading sample cost $\sum_j T_j^\star(p)$, because $\log(J/\delta)/\log(1/\delta) \to 1$. Evaluating several certificates on shared data does not impose this additive cost; their common sampling allocation is governed by the joint objective.

D Two optimization formulations for fixed-policy testing

We relate the two fixed-policy optimization problems of Ariu et al. (2026) and record their stopping implication. Assume Assumption 1, and fix $j \in [J]$, a model $p \in \mathcal{P}_{\mathrm{test}}^{1:J}$, and an allocation $\omega \in \Delta(\mathcal{S} \times \mathcal{A})$ with $\omega_{sa} > 0$ for every (s,a). Define

$$g_{p,j}(q) := V_{p,j}(\rho) V_{q,j}(\rho), \qquad h(q) := \Phi(p;\omega,q) = \sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p,q).$$

The non-convex objective (NO) and non-convex constraint (NC) formulations exchange the roles of the value product and the KL cost:

$$u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p) := \min\{g_{p,j}(q) : h(q) \le \sigma\}, \qquad \sigma \ge 0,$$
$$\sigma_{\mathrm{NC}}^{(j)}(u;\omega,p) := \inf\{h(q) : g_{p,j}(q) < u\}.$$

The NO minimum is attained because its feasible set is nonempty and compact (KL is lower semicontinuous) and policy values are continuous. At $u = 0$, the NC problem minimizes over sign-reversing alternatives:

$$\sigma_j^\star(\omega,p) := \sigma_{\mathrm{NC}}^{(j)}(0;\omega,p) = \inf_{q \in \mathrm{Alt}_j(p)} \Phi(p;\omega,q).$$

Proposition 3 (Relation between the two optimal values)**.** *Under the preceding assumptions, the optimal-value functions $u_{\mathrm{NO}}^{(j)}$ and $\sigma_{\mathrm{NC}}^{(j)}$ are inverse on the nontrivial range:*

$$\sigma_{\mathrm{NC}}^{(j)}\big(u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p);\omega,p\big)=\sigma$$

whenever

$$u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p)>\min_{q\in\mathcal{P}} g_{p,j}(q).$$

In particular,

$$\sigma_j^\star(\omega,p)=\min_{q:\, g_{p,j}(q)\le 0} h(q)=\inf\{\sigma\ge 0: u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p)\le 0\}.$$

Proof. The inverse relation and zero-level identity are Ariu et al. (2026, Proposition 2), applied to $\pi^{(j)}$. Positive row weights make zero KL radius identify p, and nontriviality places zero strictly above the minimum of $g_{p,j}$. For policies without full support, fix unused rows to p and apply the support reduction in Appendix E. Since the NO minimum is attained, $u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p)\le 0$ is equivalent to the existence of q with $g_{p,j}(q)\le 0$ and $h(q)\le\sigma$. This identifies the threshold with the closed-alternative minimum; that minimum is attained by lower semicontinuity on a compact set and is finite because a full-support strict alternative exists. □

Lemma 6 (Policywise stopping implication)**.** *If $u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p)>0$, every kernel q with $\Phi(p;\omega,q)\le\sigma$ has the same j-th sign as p.*

Proof. Every such q satisfies $V_{p,j}(\rho)V_{q,j}(\rho)\ge u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p)>0$. □

For online stopping, we certify positivity at $(\omega,p,\sigma)=(\hat{w}(t),\hat{p}_t,\sigma_t)$; localizing the entire threshold is unnecessary. Appendix E gives the certified interface.

E Using the fixed-policy solver for stopping

We instantiate the deterministic optimization guarantee of Ariu et al. (2026) for each target policy. Assume Assumption 1, fix $j\in[J]$, and use the NO objective from Appendix D:

$$u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p):=\min_{q\in\mathcal{P}:\,\Phi(p;\omega,q)\le\sigma} V_{p,j}(\rho)V_{q,j}(\rho),\qquad \Phi(p;\omega,q)=\sum_{s,a}\omega_{sa}\mathrm{KL}_{sa}(p,q).$$

Here the reference kernel $p\in\mathcal{P}$, positive allocation ω, and budget $\sigma\ge 0$ are fixed during each call; the variable is the product of transition rows q. They may be data-dependent between calls.

Reduction and hypotheses. In the reversed MDP, states are original state-action pairs (s,a), actions are next states y, and a reversed policy encodes $q(y\mid s,a)$. Its initial law is $\bar{\rho}_{sa}=\rho_s\pi^{(j)}(a\mid s)$, and its value equals $V_{q,j}(\rho)$. Thus the objective and weighted KL constraint coincide with the fixed-policy NO problem.

For a target policy without full support, restrict the reversed state space to $I_j:=\{(s,a):\pi^{(j)}(a\mid s)>0\}$. Rows outside I_j do not affect the value; fixing them to p makes their KL cost zero and cannot worsen any feasible objective. On I_j, both $\bar{\rho}$ and the row weights are positive. If normalized weights are required, let $c_j:=\sum_{(s,a)\in I_j}\omega_{sa}>0$ and use weights ω_{sa}/c_j and budget σ/c_j. Assumption 1 supplies the full-support initial state law and the nontrivial value range required by the fixed-policy result. The reference transition rows themselves may have zero entries.

Proposition 4 (Certified bounds for policywise stopping)**.** *Under Assumption 1, for every $p\in\mathcal{P}$, strictly positive allocation ω, budget $\sigma\ge 0$, and $\xi^{\mathrm{stop}}>0$, the fixed-policy solver terminates with*

$$L_j(\sigma)\le u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p)\le U_j(\sigma),\qquad U_j(\sigma)-L_j(\sigma)\le\xi^{\mathrm{stop}}.$$

Proof. If $V_{p,j}(\rho)=0$, return $[0,0]$. If $\sigma=0$, positivity of ω forces $q=p$, so return the exact value $V_{p,j}(\rho)^2$ as both endpoints. Otherwise apply Ariu et al. (2026, Theorem 2, arXiv v2) with the preceding reduction. Its output $u_j^\xi(\sigma)$ satisfies

$$u_j^\xi(\sigma)-\xi^{\mathrm{stop}}\le u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p)\le u_j^\xi(\sigma).$$

Take $L_j := u_j^{\xi} - \xi^{\mathrm{stop}}$ and $U_j := u_j^{\xi}$. □

Proposition 5 (Policywise certified stopping implication)**.** *If the certified lower bound satisfies $L_j(\sigma) > 0$, every $q \in \mathcal{P}$ with $\Phi(p;\omega,q) \le \sigma$ has the same j-th sign as p.*

Proof. For every feasible q, $V_{p,j}(\rho)V_{q,j}(\rho) \ge u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p) \ge L_j(\sigma) > 0$. □

Use along the adaptive trajectory. PT-ACE(μ) calls this interface at $(p,\omega,\sigma) = (\hat{p}_t, \hat{w}(t), \sigma_t)$ only after all rows have been observed and all empirical target values are nonzero. Coverage ensures positive weights. It stops when $\min_j L_j(\sigma_t) > 0$, certifying all signs on the same plausible set. Appendix K proves the statistical guarantee for this adaptive trajectory; only the deterministic solver certificate is imported here. Exchange additionally requires the feasible alternative kernels and objective certificates specified in Appendix F.

F Certified allocation cuts

We show how certified near-minimizing alternative kernels produce allocation cuts. This requires a feasible kernel and an objective certificate; the scalar stopping brackets in Appendix E alone do not supply them. The main upper bound assumes the kernel interface below. The reduction establishes the cut certificate without asserting a polynomial-time construction of that interface. Throughout, assume Assumption 1.

F.1 Certified alternative kernels

Exchange requires an alternative with certified near-minimal cost for the optimization over closed alternatives

$$\mathrm{Alt}_j^{\le}(p) := \{q \in \mathcal{P} : V_{p,j}(\rho)V_{q,j}(\rho) \le 0\}, \qquad F_p^{(j)}(\omega) := \inf_{q \in \mathrm{Alt}_j^{\le}(p)} \Phi(p;\omega,q).$$

Definition 2 (Certified near-minimizing alternative)**.** *Given $j \in [J]$, $p \in \mathcal{P}_{\mathrm{test}}^{1:J}$, a strictly positive allocation ω, and $\epsilon > 0$, the routine must terminate with $q_j^{\epsilon} \in \mathrm{Alt}_j^{\le}(p)$ and $\widehat{\sigma}_j$ satisfying*

$$F_p^{(j)}(\omega) \le \widehat{\sigma}_j \le F_p^{(j)}(\omega) + \epsilon, \qquad \Phi(p;\omega,q_j^{\epsilon}) \le \widehat{\sigma}_j + \epsilon.$$

Thus $\Phi(p;\omega,q_j^{\epsilon}) \le F_p^{(j)}(\omega) + 2\epsilon$.

An implementation must certify both the opposite-sign feasibility of its returned kernel and this objective bound. Keeping an uncertified primal iterate, or computing only a scalar bracket, does not meet the requirement. The analysis below depends only on the stated certificate.

F.2 From policywise alternatives to a joint allocation cut

For any reference kernel $p' \in \mathcal{P}$, let $F_{p'}^{1:J}(x) := \min_j F_{p'}^{(j)}(x)$, using closed alternatives as above. If some $V_{p',j}(\rho) = 0$, then $\mathrm{Alt}_j^{\le}(p') = \mathcal{P}$, so $q = p'$ gives $F_{p'}^{1:J} \equiv 0$. Otherwise, compute a certified alternative for each policy and select the smallest certified approximate cost, as follows.

Proposition 6 (Certified allocation cut)**.** *Algorithm 2, with policywise accuracy $\xi^{\mathrm{cut}}/4$, returns a feasible $q_\star \in \mathrm{Alt}_{j_\star}^{\le}(p')$ and a finite vector $g = (\mathrm{KL}_{sa}(p', q_\star))_{s,a}$ satisfying*

$$\Phi(p';x,q_\star) \le F_{p'}^{1:J}(x) + \xi^{\mathrm{cut}}/2. \tag{10}$$

For every $z \in \Delta(\mathcal{S}\times\mathcal{A})$, and hence every $z \in \widetilde{\Sigma}_{p'}^{\mu}$,

$$F_{p'}^{1:J}(z) - F_{p'}^{1:J}(x) \le \langle g, z - x\rangle + \xi^{\mathrm{cut}}. \tag{11}$$

Proof. The zero-value branch has $q_\star = p'$, $g = 0$, and $F_{p'}^{1:J} \equiv 0$. Otherwise choose $j^\circ \in \arg\min_j F_{p'}^{(j)}(x)$. The certificates and the choice of $j_\star$ give

$$\Phi(p';x,q_\star) \le \widehat{\sigma}_{j_\star} + \xi^{\mathrm{cut}}/4 \le \widehat{\sigma}_{j^\circ} + \xi^{\mathrm{cut}}/4 \le F_{p'}^{1:J}(x) + \xi^{\mathrm{cut}}/2.$$

Algorithm 2 Certified allocation cut

Require: model $p' \in \mathcal{P}$, query $x \in K_\mu$, accuracy $\xi^{\mathrm{cut}} > 0$
if some $j \in [J]$ satisfies $V_{p',j}(\rho) = 0$ **then**
  return $(j, p', 0)$ for any such index
end if
for $j = 1, \dots, J$ **do**
  Obtain $(q_j, \widehat{\sigma}_j)$ from Definition 2 at (j, p', x), with accuracy $\epsilon = \xi^{\mathrm{cut}}/4$
end for
$j_\star \in \arg\min_j \widehat{\sigma}_j, \quad q_\star \leftarrow q_{j_\star}$
$g_{sa} \leftarrow \mathrm{KL}_{sa}(p', q_\star)$ for every (s, a)
return $(j_\star, q_\star, g)$

Because $x \in K_\mu$ and all KL terms are nonnegative, finite weighted cost implies that every coordinate of g is finite. Feasibility gives

$$F_{p'}^{1:J}(z) \le \Phi(p'; z, q_\star) = \Phi(p'; x, q_\star) + \langle g, z - x \rangle.$$

Combining these bounds proves the cut inequality, even with error $\xi^{\mathrm{cut}}/2$. The selected policy need not minimize the exact costs. □

In PT-ACE(μ), substitute $(p', x, \xi^{\mathrm{cut}}) = (\bar{p}_t, x_t, \xi_t^{\mathrm{cut}})$. The objective certificate and cut inequality are then exactly those used in the exchange analysis of Appendix H.

G Concentration of the empirical kernels

We prove Proposition 7, the kernel-concentration bound used in the exchange analysis: for $\tau_T := \lceil T^{1/4} \rceil$, every fixed $\zeta > 0$, and all sufficiently large T,

$$\mathcal{C}_T^{\mathrm{ker}}(\zeta) := \bigcap_{t=\tau_T}^{T} \{\|\hat{p}_t - p\|_\infty \le \zeta\}, \qquad \mathbb{P}\big((\mathcal{C}_T^{\mathrm{ker}}(\zeta))^c\big) \le B_{\mu,p,\zeta} T e^{-C_{\mu,p,\zeta}\tau_T}.$$

Lemmas 7 and 8 turn the action floor into a common minorization. Lemma 9 then gives linear visit counts with exponential tails. Combining this count bound with transition martingales yields Lemma 10; a union bound over the later rounds proves the proposition.

Throughout, assume Assumption 2 and fix an algorithmic floor $0 < \mu \le 1/(|\mathcal{S}||\mathcal{A}|)$. For analysis, continue the sampling rule after stopping. Let $Z_t = (S_t, A_t)$ and let $\mathcal{F}_t$ be the history before drawing S_{t+1}, including Z_t and the sampling policy χ_t. Under the one-step update convention,

$$S_{t+1} \sim p(\cdot \mid S_t, A_t), \qquad A_{t+1} \sim \chi_t(\cdot \mid S_{t+1}), \qquad P_t\big((s,a),(s',a')\big) := p(s' \mid s,a)\chi_t(a' \mid s').$$

Write $P^u := P_p^{\mathrm{Unif}_{\mathcal{A}}}$, $d := |\mathcal{S}||\mathcal{A}|$, and

$$N_{sa}(t) := \sum_{u<t} \mathbf{1}\{Z_u = (s,a)\}, \qquad \|\hat{p}_t - p\|_\infty := \max_{s,a} \|\hat{p}_t(\cdot \mid s,a) - p(\cdot \mid s,a)\|_1.$$

Unobserved empirical rows may be defined arbitrarily.

G.1 The action floor and common minorization

Lemma 7 (Uniform action floor)**.** *Every $t \ge 0$, $\chi_t(a \mid s) \ge \mu$ for all (s,a), and $P_t \ge \lambda_\mu P^u$ entrywise, where $\lambda_\mu := |\mathcal{A}|\mu$.*

Proof. Since $\bar{x}_t \in K_\mu$ and $\sum_b \bar{x}_t(s,b) \le 1$, $\chi_t(a \mid s) = \bar{x}_t(s,a)/\sum_b \bar{x}_t(s,b) \ge \mu$. Substituting this bound into P_t gives the matrix inequality. □

Lemma 8 (Common minorization for fixed and adaptive kernels)**.** *There are $m_\mu \ge 1$, $\varepsilon_\mu \in (0,1)$, and a full-support probability distribution ν_μ, depending only on (μ, p), such that any stochastic matrices $Q_1, \dots, Q_{m_\mu}$ satisfying $Q_k \ge \lambda_\mu P^u$ obey*

$$Q_1 \cdots Q_{m_\mu}(z, \cdot) \ge \varepsilon_\mu \nu_\mu(\cdot).$$

In particular, this applies to each frozen power $P_t^{m_\mu}$ and to every realized product of consecutive P_t's. The adaptive trajectory also satisfies the conditional bound

$$\mathbb{P}(Z_{i+m_\mu} \in \cdot \mid \mathcal{F}_i) \geq \varepsilon_\mu \nu_\mu(\cdot) \qquad (i \geq 0).$$

Proof. Lemma 7 supplies the domination $P_t \geq \lambda_\mu P^u$ needed below. The finite irreducible aperiodic matrix P^u has a strictly positive power. Choose m_μ with $r := \min_{z,z'}(P^u)^{m_\mu}(z,z') > 0$, let ν_μ be uniform on the d state-action pairs, and set

$$\varepsilon_\mu := \tfrac{1}{2} d \lambda_\mu^{m_\mu} r \in (0,1).$$

Nonnegative matrix multiplication gives $Q_1 \cdots Q_{m_\mu} \geq \lambda_\mu^{m_\mu}(P^u)^{m_\mu} \geq \varepsilon_\mu \mathbf{1}\nu_\mu$.

For the adaptive trajectory, use conditional expectations rather than interpreting a product of future random matrices as its transition law. For every deterministic nonnegative function f,

$$\mathbb{E}[f(Z_{t+1}) \mid \mathcal{F}_t] = P_t f(Z_t) \geq \lambda_\mu P^u f(Z_t).$$

Iterating the tower property with the deterministic functions $f, P^u f, \ldots, (P^u)^{m_\mu - 1} f$ yields

$$\mathbb{E}[f(Z_{i+m_\mu}) \mid \mathcal{F}_i] \geq \lambda_\mu^{m_\mu}(P^u)^{m_\mu} f(Z_i) \geq \varepsilon_\mu \nu_\mu(f),$$

which proves the conditional bound. □

G.2 Linear visit counts with exponential tails

Lemma 9 (Visit-count lower bound at a fixed time)**.** *There are $c_{\mu,p}, C_{\mu,p} > 0$ such that, for every (s,a) and all sufficiently large t,*

$$\mathbb{P}\big(N_{sa}(t) \leq c_{\mu,p} t\big) \leq e^{-C_{\mu,p} t}, \qquad \mathbb{P}\big(\min_{s,a} N_{sa}(t) \leq c_{\mu,p} t\big) \leq d e^{-C_{\mu,p} t}.$$

Proof. Put $m = m_\mu$ and $q_\star := \varepsilon_\mu \min_z \nu_\mu(z) > 0$. For a fixed pair (s,a), let $Y_k := \mathbf{1}\{Z_{km} = (s,a)\}$ and $\mathcal{G}_k := \mathcal{F}_{km}$. The conditional minorization in Lemma 8 gives $\mathbb{E}[Y_k \mid \mathcal{G}_{k-1}] \geq q_\star$. Thus $M_n := \sum_{k=1}^n (Y_k - \mathbb{E}[Y_k \mid \mathcal{G}_{k-1}])$ is a martingale with increments bounded by one, and Azuma–Hoeffding gives

$$\mathbb{P}\left(\sum_{k=1}^n Y_k \leq n q_\star / 2\right) \leq \mathbb{P}(M_n \leq -n q_\star/2) \leq e^{-n q_\star^2/8}.$$

For $n_t := \lfloor (t-1)/m \rfloor$, all these endpoints are counted in $N_{sa}(t)$, so $N_{sa}(t) \geq \sum_{k=1}^{n_t} Y_k$. For large t, $n_t \geq t/(2m)$, giving the claim with

$$c_{\mu,p} := \frac{q_\star}{4m}, \qquad C_{\mu,p} := \frac{q_\star^2}{16m}.$$

The second bound is a union bound over the d pairs. □

G.3 Empirical-kernel concentration

Lemma 10 (Kernel concentration at a fixed time)**.** *For every $\zeta > 0$, there are $B_{\mu,p,\zeta}, C_{\mu,p,\zeta} > 0$ such that, for all sufficiently large t,*

$$\mathbb{P}(\|\hat{p}_t - p\|_\infty > \zeta) \leq B_{\mu,p,\zeta} e^{-C_{\mu,p,\zeta} t}.$$

Proof. For each (s,a,y), define

$$M_{sa,y}(t) := \sum_{u<t} \mathbf{1}\{Z_u = (s,a)\}\big(\mathbf{1}\{S_{u+1} = y\} - p(y \mid s,a)\big).$$

Conditional on $\mathcal{F}_u$, each summand has mean zero and absolute value at most one. Hence $M_{sa,y}$ is an $(\mathcal{F}_t)$-martingale. Whenever $N_{sa}(t) > 0$, the empirical-row definition gives

$$M_{sa,y}(t) = N_{sa}(t)\big(\hat{p}_t(y \mid s,a) - p(y \mid s,a)\big).$$

Let $c = c_{\mu,p}$ and $C = C_{\mu,p}$ be the constants from Lemma 9. If $N_{sa}(t) > ct$ but the row's ℓ_1 error exceeds ζ, then $|M_{sa,y}(t)| > c\zeta t/|\mathcal{S}|$ for some y. Lemma 9, Azuma–Hoeffding, and a union bound therefore give

$$\begin{aligned}\mathbb{P}(\|\hat{p}_t - p\|_\infty > \zeta) &\le \sum_{s,a} \mathbb{P}(N_{sa}(t) \le ct) + \sum_{s,a,y} \mathbb{P}\big(|M_{sa,y}(t)| > c\zeta t/|\mathcal{S}|\big) \\ &\le d\left[e^{-Ct} + 2|\mathcal{S}| \exp\left(-\frac{c^2\zeta^2 t}{2|\mathcal{S}|^2}\right)\right].\end{aligned}$$

Thus take $B_{\mu,p,\zeta} := d(1 + 2|\mathcal{S}|)$ and $C_{\mu,p,\zeta} := \min\{C, c^2\zeta^2/(2|\mathcal{S}|^2)\}$. This argument uses unconditional martingale bounds; it never conditions on a realized adaptive count $N_{sa}(t)$. □

Proposition 7 (Kernel concentration over later rounds)**.** *For every $\zeta > 0$ and all sufficiently large T,*

$$\mathbb{P}\big((\mathcal{C}_T^{\ker}(\zeta))^c\big) \le B_{\mu,p,\zeta} T e^{-C_{\mu,p,\zeta}\tau_T}.$$

Proof. For fixed $\zeta > 0$, Lemma 10 provides constants $B_{\mu,p,\zeta}, C_{\mu,p,\zeta} > 0$ and a threshold t_0 such that its bound holds for every $t \ge t_0$. Choose T large enough that $\tau_T \ge t_0$. Applying that lemma at each $t \in [\tau_T, T]$ and taking a union bound gives

$$\mathbb{P}\big((\mathcal{C}_T^{\ker}(\zeta))^c\big) \le \sum_{t=\tau_T}^{T} B_{\mu,p,\zeta} e^{-C_{\mu,p,\zeta}t} \le B_{\mu,p,\zeta} T e^{-C_{\mu,p,\zeta}\tau_T}.$$

□

H Exchange analysis for PT-ACE(μ)

The main result is Proposition 8: the learned allocation approaches the constrained information benchmark. It establishes the exchange step in the proof sketch of Section 3.6. We state it first, then develop its proof. Lemma 16 supplies the bounded-cut argument. A companion flow bound, Proposition 9, provides the other input to the navigation analysis.

Throughout the exchange analysis, assume Assumption 1, fix $p \in \mathcal{P}_{\mathrm{adm}}^{1:J}$ and a floor $\mu > 0$ satisfying $\Sigma_p \cap K_{2\mu} \neq \varnothing$, and use the certified cuts, exact occupancy updates, and tolerances in Theorem 2. Write $\|p' - p\|_\infty := \max_{s,a} \|p'(\cdot \mid s,a) - p(\cdot \mid s,a)\|_1$.

Set $\tau_T := \lceil T^{1/4} \rceil$ and $\vartheta_T := \lceil T^{3/4} \rceil$, and recall

$$\mathcal{C}_T^{\ker}(\zeta) := \bigcap_{t=\tau_T}^{T} \{\|\hat{p}_t - p\|_\infty \le \zeta\}.$$

Proposition 8 (Objective bound over later rounds)**.** *Under the exchange hypotheses above, for each $\zeta > 0$ there is a deterministic remainder $r_{\mathrm{ex}}^{\mathrm{tail}}(T,\mu,\zeta) \ge 0$ such that, on $\mathcal{C}_T^{\ker}(\zeta)$, for all sufficiently large T,*

$$\inf_{\vartheta_T \le t \le T} F_p^{1:J}(\bar{x}_t) \ge \frac{1}{T_{2\mu}^\star(p)} - r_{\mathrm{ex}}^{\mathrm{tail}}(T,\mu,\zeta), \qquad \lim_{\zeta \downarrow 0} \limsup_{T\to\infty} r_{\mathrm{ex}}^{\mathrm{tail}}(T,\mu,\zeta) = 0.$$

Proof roadmap. Choose a maximizing occupancy in $\Sigma_p \cap K_{2\mu}$ and transport its policy to the empirical kernels. Lemma 13 keeps this comparator feasible and bounds its movement. Lemma 16 controls the cuts, and Lemma 17 gives the one-step KL inequality. Summing it bounds the weighted objective deficit; the continuity estimate in Lemma 15 and the averaging bounds then transfer the result to $F_p^{1:J}(\bar{x}_t)$. Section H.5 assembles these ingredients and proves the companion flow bound.

H.1 Objective and exchange update

Recall the closed-alternative objectives

$$\begin{aligned}&\mathrm{Alt}_j^{\le}(p') := \{q \in \mathcal{P} : V_{p',j}(\rho) V_{q,j}(\rho) \le 0\}, \\ &F_{p'}^{(j)}(w) := \min_{q \in \mathrm{Alt}_j^{\le}(p')} \Phi(p'; w, q), \qquad \Phi(p'; w, q) := \sum_{s,a} w_{sa} \mathrm{KL}_{sa}(p', q), \\ &F_{p'}^{1:J}(w) := \min_{j \in [J]} F_{p'}^{(j)}(w).\end{aligned}$$

These functions are nonnegative and concave in w. For positive allocations and $p' \in \mathcal{P}_{\text{test}}^{1:J}$, Appendix D identifies the closed-alternative minimum with the strict-alternative infimum.

Use the notation

$$\bar{p}_t = (1-\theta_t)\hat{p}_t + \theta_t \text{Unif}_{\mathcal{S}}, \quad \theta_t = (t \vee 1)^{-1/4}, \quad D_t := \Sigma_{\bar{p}_t} \cap K_\mu, \quad F_t := F_{\bar{p}_t}^{1:J}.$$

A round is *active* when both D_t and D_{t+1} are nonempty. On such a round, x_t denotes the query center after any feasibility reset into D_t. The cut certificate and update are

$$g_t(s,a) = \text{KL}_{sa}(\bar{p}_t, q_t), \qquad q_t \in \text{Alt}_{\bar{J}_t}^{\leq}(\bar{p}_t), \qquad \langle g_t, x_t \rangle \leq F_t(x_t) + \xi_t^{\text{cut}},$$
$$G_{t+1} = G_t + \|g_t\|_\infty^2, \qquad \eta_t = G_{t+1}^{-1/2}, \qquad G_0 > 0,$$
$$y_{t+1} \in \arg\max_{x \in D_t} \{\langle g_t, x \rangle - \eta_t^{-1} \text{KL}(x \| x_t)\}, \qquad x_{t+1} \in \arg\min_{x \in D_{t+1}} \text{KL}(x \| y_{t+1}).$$

Both kernels are available at this update, after the navigation transition. On a skipped round, set $g_t = 0$, retain $x_{t+1} = x_t$, and use the same accumulator and stepsize rules. The average always uses the corrected or retained center:

$$\bar{x}_t = \frac{\sum_{u=0}^{t-1} \eta_u x_{u+1}}{\sum_{u=0}^{t-1} \eta_u}.$$

All centers, predictors, and averages belong to K_μ.

H.2 A feasible comparator on the changing domains

Write $\alpha_p := \alpha_p^{\text{Unif}_{\mathcal{A}}}$ and $\underline{\alpha}_p := \min_{s,a} \alpha_p(s,a) > 0$.

Lemma 11 (The floor-constrained benchmark)**.** *Whenever $\Sigma_p \cap K_{2\mu} \neq \varnothing$, the maximum in* (3) *is attained at some $w^{2\mu}$, with*

$$F_p^{1:J}(w^{2\mu}) = \frac{1}{T_{2\mu}^\star(p)}, \qquad 0 < T^\star(p) \leq T_{2\mu}^\star(p) < \infty.$$

Moreover, for $0 < 2\mu < \underline{\alpha}_p$, this set is nonempty and

$$T_{2\mu}^\star(p) \leq \frac{T^\star(p)}{1 - 2\mu/\underline{\alpha}_p}.$$

Consequently, $T_{2\mu}^\star(p) \to T^\star(p)$ as $\mu \downarrow 0$.

Proof. If the constrained set is nonempty, compactness and continuity on positive allocations (Lemma 15) give a maximizer $w^{2\mu}$. Its objective is finite because nontriviality supplies full-support strict alternatives. It is positive because the closed alternative sets are compact and exclude p, while positive weights make the weighted KL vanish only at $q = p$.

To compare with the unconstrained benchmark, write

$$F_{\text{str}}(w) := \inf_{q \in \text{Alt}_{1:J}(p)} \Phi(p; w, q)$$

for the strict-alternative objective of Section 2.2. It is nonnegative and concave. Lemma 5 expresses it as an infimum of finite linear functions, so it is upper semicontinuous and attains its finite maximum $1/T^\star(p)$ on Σ_p. At positive allocations it equals the closed-alternative objective by Appendix D. Set inclusion therefore gives $0 < T^\star(p) \leq T_{2\mu}^\star(p) < \infty$.

Now suppose $0 < 2\mu < \underline{\alpha}_p$. Choose a maximizer $w^\star$ of F_{str} on Σ_p, and put $\theta_\mu := 2\mu/\underline{\alpha}_p < 1$. The positive mixture $\widetilde{w} = (1-\theta_\mu)w^\star + \theta_\mu \alpha_p$ belongs to $\Sigma_p \cap K_{2\mu}$. Concavity of the strict objective, followed by its equality with the closed objective at $\widetilde{w}$, yields

$$F_p^{1:J}(w^{2\mu}) \geq F_p^{1:J}(\widetilde{w}) = F_{\text{str}}(\widetilde{w}) \geq \frac{1-\theta_\mu}{T^\star(p)}.$$

Taking reciprocals proves the small-floor bound and the stated limit. □

Define the fixed policy and its transported occupancies by

$$\chi^\mu(a \mid s) := \frac{w_{sa}^{2\mu}}{\sum_b w_{sb}^{2\mu}}, \qquad w_t^\mu := \alpha_{\bar{p}_t}^{\chi^\mu}.$$

Then $\chi^\mu(a \mid s) \geq 2\mu$ and $\alpha_p^{\chi^\mu} = w^{2\mu}$. The following perturbation estimate controls both feasibility and movement of w_t^μ.

Lemma 12 (Occupancy perturbation under a fixed policy)**.** *Under Assumption 2, for each action floor $\underline{\chi} > 0$ there are $\zeta_\chi > 0$ and $L_{\mathrm{occ}}(\underline{\chi}, p) < \infty$ such that, uniformly over policies $\chi(a \mid s) \geq \underline{\chi}$,*

$$\|\alpha_{p'}^\chi - \alpha_{p''}^\chi\|_1 \leq L_{\mathrm{occ}}(\underline{\chi}, p)\|p' - p''\|_\infty \quad \textit{if } \|p' - p\|_\infty, \|p'' - p\|_\infty \leq \zeta_\chi.$$

Proof. A power of the uniform-policy state-action kernel at p has all entries positive. This remains true, with a common lower bound, near p. Since $P_{p'}^\chi \geq |\mathcal{A}|\underline{\chi} P_{p'}^{\mathrm{Unif}_\mathcal{A}}$, these kernels have a common Doeblin minorization and a uniform geometric contraction on zero-mass row vectors. Put $P' = P_{p'}^\chi$, $P'' = P_{p''}^\chi$, and $\alpha' = \alpha_{p'}^\chi$, $\alpha'' = \alpha_{p''}^\chi$. Iterating $\alpha' - \alpha'' = (\alpha' - \alpha'')P'' + \alpha'(P' - P'')$ gives

$$\alpha' - \alpha'' = \sum_{k \geq 0} \alpha'(P' - P'')(P'')^k.$$

The common contraction bounds this series by a constant times $\max_z \|P'(z, \cdot) - P''(z, \cdot)\|_1 \leq \|p' - p''\|_\infty$. □

To apply Lemma 12 to the transported comparator $w_t^\mu = \alpha_{\bar{p}_t}^{\chi^\mu}$, we need to control the stabilized kernel $\bar{p}_t$, whereas $\mathcal{C}_T^{\mathrm{ker}}(\zeta)$ controls the raw estimate $\hat{p}_t$. By the smoothing definition in Section H.1, $\bar{p}_t - p = (1 - \theta_t)(\hat{p}_t - p) + \theta_t(\mathrm{Unif}_\mathcal{S} - p)$. Since two probability rows have ℓ_1 distance at most two, on $\mathcal{C}_T^{\mathrm{ker}}(\zeta)$,

$$\|\bar{p}_t - p\|_\infty \leq (1 - \theta_t)\zeta + 2\theta_t \leq \zeta + 2\tau_T^{-1/4} =: d_T(\zeta), \qquad \tau_T \leq t \leq T. \tag{12}$$

Thus small ζ and large T place every $\bar{p}_t$ in the perturbation neighborhood of that lemma. The next lemma uses this bound to keep w_t^μ within the floor slack between $K_{2\mu}$ and K_μ, ensuring feasibility throughout the tail window.

Lemma 13 (Feasibility and movement of the comparator)**.** *For sufficiently small $\zeta > 0$ and sufficiently large T, on $\mathcal{C}_T^{\mathrm{ker}}(\zeta)$, with $L := L_{\mathrm{occ}}(2\mu, p)$,*

$$w_t^\mu \in D_t, \qquad \|w_t^\mu - w^{2\mu}\|_1 \leq L d_T(\zeta) \quad (\tau_T \leq t \leq T),$$
$$\sum_{t=\tau_T}^{T-1} \|w_{t+1}^\mu - w_t^\mu\|_1 \leq C_{\mathrm{drift}}(\mu, p) \log(eT).$$

In particular, every round in $[\tau_T, T - 1]$ is active.

Proof. Apply Lemma 12 with $\chi = \chi^\mu$, $p' = \bar{p}_t$, and $p'' = p$, using (12). Since $\alpha_p^{\chi^\mu} = w^{2\mu}$, this gives the proximity bound. Choose ζ small and T large so that $L d_T(\zeta) \leq \mu$ and the kernels remain in its neighborhood. Since $w^{2\mu} \in K_{2\mu}$, this gives $w_t^\mu \in K_\mu$; stationarity gives $w_t^\mu \in \Sigma_{\bar{p}_t}$. For the movement bound, apply Lemma 12 to consecutive stabilized kernels to obtain

$$\|w_{t+1}^\mu - w_t^\mu\|_1 \leq L\big(\|\hat{p}_{t+1} - \hat{p}_t\|_\infty + 2|\theta_{t+1} - \theta_t|\big).$$

Only the observed row changes, by at most $2/(N_{S_t A_t}(t+1) \vee 1)$. Grouping these increments by row bounds their sum through time T by $2|\mathcal{S}||\mathcal{A}|(1 + \log T)$, while the total variation of θ_t is at most one. This proves the claim without a visit-count event. □

H.3 A common continuity lemma for the KL objectives

Lemma 14 (Continuity of the policy value in the kernel)**.** *For every fixed policy χ, $p' \mapsto V_{p'}^\chi(\rho)$ is continuous on $\mathcal{P}$.*

Proof. Use $V^{\chi}_{p'}(\rho) = \rho^{\top}(I - \gamma P^{\chi}_{p'})^{-1} r^{\chi}$. The transition matrix is affine in p', and the inverse exists for every stochastic matrix because $\gamma < 1$. □

Lemma 15 (Objective perturbation on positive allocations)**.** *Assume Assumption 1 and $p \in \mathcal{P}^{1:J}_{\mathrm{test}}$. For every fixed allocation floor $\kappa > 0$ with $K_\kappa \neq \varnothing$, there is a closed neighborhood $\mathcal{P}_{\zeta_2}(p) := \{p' \in \mathcal{P} : \|p' - p\|_\infty \le \zeta_2\}$ on which every $F^{(j)}_{p'}(w)$, and hence $F^{1:J}_{p'}(w)$, is finite and jointly continuous in $(p', w) \in \mathcal{P}_{\zeta_2}(p) \times K_\kappa$. For $\kappa = \mu$, the exchange analysis therefore has a deterministic remainder $\varepsilon_F(\mu, \zeta, T) \ge 0$ such that, on $\mathcal{C}^{\mathrm{ker}}_T(\zeta)$, for sufficiently small ζ, large T, and every $\tau_T \le t \le T$,*

$$|F^{1:J}_{\bar{p}_t}(w^{\mu}_t) - F^{1:J}_{p}(w^{2\mu})| \le \varepsilon_F(\mu, \zeta, T),$$
$$\sup_{x \in K_\mu} |F^{1:J}_{\bar{p}_t}(x) - F^{1:J}_{p}(x)| \le \varepsilon_F(\mu, \zeta, T), \qquad \lim_{\zeta \downarrow 0} \limsup_{T \to \infty} \varepsilon_F(\mu, \zeta, T) = 0.$$

Proof. Value continuity and the nonzero target values at p let us choose the neighborhood so that all signs are fixed. For each j, put $b_j := \mathrm{sign}(V_{p,j}(\rho))$. The closed alternative set is then the fixed compact set $\Gamma_j := \{q \in \mathcal{P} : b_j V_{q,j}(\rho) \le 0\}$. Nontriviality supplies a strict opposite-sign kernel; mixing it slightly with the rowwise uniform kernel preserves its sign and gives full support. Thus the KL minimum is finite and is attained by lower semicontinuity.

For lower semicontinuity, let $(p_n, w_n) \to (p_0, w_0)$ in the stated set. Choose a subsequence attaining the liminf of the optimal values, and then a convergent subsequence of minimizers $q_n \to q_\infty \in \Gamma_j$. Lower semicontinuity of KL, with weights bounded below by κ, yields

$$F^{(j)}_{p_0}(w_0) \le \Phi(p_0; w_0, q_\infty) \le \liminf_n \Phi(p_n; w_n, q_n) = \liminf_n F^{(j)}_{p_n}(w_n).$$

For upper semicontinuity, Appendix D allows a strict alternative q_ε with finite cost at most $F^{(j)}_{p_0}(w_0) + \varepsilon$. Let $q_{\varepsilon,h} := (1-h)q_\varepsilon + h\mathrm{Unif}_{\mathcal{S}}$. For sufficiently small $h > 0$, it remains strict and has full row support. Moreover,

$$\Phi(p_0; w_0, q_{\varepsilon,h}) \le \Phi(p_0; w_0, q_\varepsilon) - \log(1-h).$$

Choose h so that $-\log(1-h) \le \varepsilon$, and keep it fixed as $n \to \infty$. Its KL cost is continuous in (p', w), including when p' has zero entries. Since it belongs to the fixed set Γ_j,

$$\limsup_n F^{(j)}_{p_n}(w_n) \le \Phi(p_0; w_0, q_{\varepsilon,h}) \le F^{(j)}_{p_0}(w_0) + 2\varepsilon.$$

Letting $\varepsilon \downarrow 0$ proves continuity. A finite minimum preserves continuity, and compactness makes it uniform. No continuity of KL on the full product of simplices is used.

For the exchange bounds, let ω be a common nondecreasing modulus on a sufficiently small compact neighborhood times K_μ. Lemma 13 bounds the joint distance of $(\bar{p}_t, w^{\mu}_t)$ from $(p, w^{2\mu})$ by $(1 + L)d_T(\zeta)$. Thus one may take $\varepsilon_F(\mu, \zeta, T) = \omega((1+L)d_T(\zeta))$. □

H.4 The one-step inequality and its cumulative error

Lemma 16 (Cut inequalities)**.** *On every active round,*

$$F_t(w) - F_t(x_t) \le \langle g_t, w - x_t \rangle + \xi^{\mathrm{cut}}_t \qquad (w \in D_t).$$

There is also a global constant $L_\mu < \infty$ such that $\|g_t\|_\infty \le L_\mu$ for all t, including skipped rounds.

Proof. The first inequality follows by subtracting the cut certificate from $F_t(w) \le \Phi(\bar{p}_t; w, q_t) = \langle g_t, w \rangle$. For the bound, choose full-support kernels q^+_j, q^-_j with opposite strict signs for every j, as in the preceding continuity proof. Then

$$C^{\circ} := \max_{j,\, b \in \{+,-\}} \max_{p' \in \mathcal{P},\, s,a} \mathrm{KL}_{sa}(p', q^b_j) < \infty, \qquad F^{1:J}_{p'}(w) \le C^{\circ}.$$

Since $g_t \ge 0$ and $x_t \in K_\mu$, on active rounds

$$\mu \|g_t\|_\infty \le \langle g_t, x_t \rangle \le C^{\circ} + \sup_u \xi^{\mathrm{cut}}_u.$$

On skipped rounds $g_t = 0$. Increase L_μ, if needed, so that $L^2_\mu \ge G_0$. □

Lemma 17 (One-step inequality with changing feasible sets)**.** *On an active round, for $w_t \in D_t$ and $w_{t+1} \in D_{t+1}$,*

$$\eta_t\big(F_t(w_t) - F_t(x_t)\big) \le \mathrm{KL}(w_t\|x_t) - \mathrm{KL}(w_{t+1}\|x_{t+1}) + C_\mu^{\mathrm{KL}}\|w_{t+1} - w_t\|_1 + \tfrac{1}{2}\eta_t^2\|g_t\|_\infty^2 + \eta_t\xi_t^{\mathrm{cut}},$$

where $C_\mu^{\mathrm{KL}} := 1 + \log(1/\mu)$ is sufficient.

Proof. Predictor optimality and the KL three-point identity give, for $w \in D_t$,

$$\eta_t\langle g_t, w - y_{t+1}\rangle \le \mathrm{KL}(w\|x_t) - \mathrm{KL}(w\|y_{t+1}) - \mathrm{KL}(y_{t+1}\|x_t).$$

Pinsker's inequality and $ab - b^2/2 \le a^2/2$, applied with $a = \eta_t\|g_t\|_\infty$ and $b = \|y_{t+1} - x_t\|_1$, then yield

$$\eta_t\big(F_t(w) - F_t(x_t)\big) \le \mathrm{KL}(w\|x_t) - \mathrm{KL}(w\|y_{t+1}) + \tfrac{1}{2}\eta_t^2\|g_t\|_\infty^2 + \eta_t\xi_t^{\mathrm{cut}}. \tag{13}$$

On K_μ, the derivative of $u \mapsto \mathrm{KL}(u\|y)$ has coordinates $1 + \log(u_{sa}/y_{sa})$ of absolute value at most C_μ^{KL}. The correction's KL Pythagorean inequality therefore implies

$$\mathrm{KL}(w_t\|y_{t+1}) \ge \mathrm{KL}(w_{t+1}\|y_{t+1}) - C_\mu^{\mathrm{KL}}\|w_{t+1} - w_t\|_1 \ge \mathrm{KL}(w_{t+1}\|x_{t+1}) - C_\mu^{\mathrm{KL}}\|w_{t+1} - w_t\|_1. \tag{14}$$

Substitution proves the claim. □

Weights and averaging. For $0 \le m < t$, put $W_{m:t} := \sum_{u=m}^{t-1}\eta_u$. The global cut bound and the accumulator identity give

$$G_{u+1} \le L_\mu^2(u+2), \qquad \frac{1}{L_\mu\sqrt{u+2}} \le \eta_u \le G_0^{-1/2}, \qquad W_{m:t} \ge \frac{t-m}{L_\mu\sqrt{t+1}},$$

$$\sum_{u=m}^{t-1}\eta_u^2\|g_u\|_\infty^2 = \sum_{u=m}^{t-1}\frac{G_{u+1} - G_u}{G_{u+1}} \le \log\frac{G_t}{G_m}.$$

For $m = \tau_T$ and $\vartheta_T \le t \le T$, these imply uniformly

$$W_{m:t} \ge C^{-1}T^{3/8}, \qquad A_{0:m} := \sum_{u<m}\eta_u \le mG_0^{-1/2}.$$

Here and below, C depends only on (μ, p) and the fixed numerical choices, and may increase between displays.

On the active interval $[m, T-1]$, only its first round can require a feasibility reset: each subsequent center is the previous correction. With x_m denoting the reset query center if necessary, define

$$\bar z_{m:t} := W_{m:t}^{-1}\sum_{u=m}^{t-1}\eta_u x_u, \qquad \bar x_{m:t} := W_{m:t}^{-1}\sum_{u=m}^{t-1}\eta_u x_{u+1}.$$

The weights are nonincreasing, so summation by parts gives

$$\left\|\sum_{u=m}^{t-1}\eta_u(x_{u+1} - x_u)\right\|_1 \le \eta_m + \sum_{u=m+1}^{t-1}(\eta_{u-1} - \eta_u) + \eta_{t-1} = 2\eta_m,$$

$$\|\bar x_t - \bar z_{m:t}\|_1 \le \frac{2A_{0:m} + 2\eta_m}{W_{m:t}} \le CT^{-1/8}.$$

The prefix average in the second line retains the centers actually recorded before the entrance reset. Thus the reset costs only the initial KL term and does not change earlier contributions to $\bar x_t$.

H.5 Completing the objective bound and controlling flow

Proof of Proposition 8. Work on the kernel event with ζ small and T large, and set $m = \tau_T$. Lemma 13 makes every round in $[m, T-1]$ active and w_u^μ feasible. Apply Lemma 17 with $w_u = w_u^\mu$ and sum from $u = m$ to $t-1$. The KL terms telescope, with initial term at most $\log(1/\mu)$; comparator

movement and the squared-step sum each cost $O(\log(eT))$. Transferring both objective terms with Lemma 15 gives

$$\sum_{u=m}^{t-1} \eta_u \big(F_p^{1:J}(w^{2\mu}) - F_p^{1:J}(x_u)\big) \le C\log(eT) + W_{m:t}\left(\sup_{u\ge m} \xi_u^{\mathrm{cut}} + 2\varepsilon_F(\mu,\zeta,T)\right).$$

Divide by $W_{m:t}$, use concavity, and pass from $\bar{z}_{m:t}$ to $\bar{x}_t$ with the averaging bound in Section H.4. If ω_μ is a nondecreasing continuity modulus for $F_p^{1:J}$ on K_μ, the result follows from $F_p^{1:J}(w^{2\mu}) = 1/T_{2\mu}^\star(p)$ with

$$r_{\mathrm{ex}}^{\mathrm{tail}}(T,\mu,\zeta) := \frac{C\log(eT)}{T^{3/8}} + \sup_{u\ge\tau_T} \xi_u^{\mathrm{cut}} + 2\varepsilon_F(\mu,\zeta,T) + \omega_\mu(CT^{-1/8}).$$

Each term vanishes in the stated order of limits. For ζ outside the small neighborhood, use the constant remainder $1/T_{2\mu}^\star(p)$, which is valid by nonnegativity and does not affect the limit. □

Proposition 9 (Flow-balance error over later rounds)**.** *Under the same hypotheses, for each $\zeta > 0$ there is a deterministic $r_{\mathrm{flow}}(T,\zeta) \ge 0$ such that, on $\mathcal{C}_T^{\mathrm{ker}}(\zeta)$, for all sufficiently large T,*

$$\sup_{\vartheta_T \le t \le T} \sum_s \left| \sum_a \bar{x}_t(s,a) - \sum_{y,b} \bar{x}_t(y,b) p(s \mid y,b) \right| \le r_{\mathrm{flow}}(T,\zeta), \qquad \lim_{\zeta\downarrow 0} \limsup_{T\to\infty} r_{\mathrm{flow}}(T,\zeta) = 0.$$

Proof. Let $B_p x$ denote the vector inside the absolute values. On the active interval, $x_{u+1} \in \Sigma_{\bar{p}_{u+1}}$, hence

$$\|B_p x_{u+1}\|_1 = \left\| \sum_{y,b} x_{u+1}(y,b)(\bar{p}_{u+1} - p)(\cdot \mid y,b) \right\|_1 \le d_T(\zeta).$$

For every simplex allocation, $\|B_p x\|_1 \le 2$. Averaging the active suffix and the recorded prefix separately therefore gives, uniformly for $\vartheta_T \le t \le T$,

$$\|B_p \bar{x}_t\|_1 \le d_T(\zeta) + \frac{2A_{0:\tau_T}}{W_{\tau_T:t}} \le \zeta + 2\tau_T^{-1/4} + CT^{-1/8}.$$

This is a valid remainder for small ζ; take $r_{\mathrm{flow}} = 2$ otherwise. Its stated limit is immediate. □

I Navigation

Our final goal is Proposition 11: transfer the exchange objective bound from the learned allocations $\bar{x}_t$ to the observed frequencies $\hat{w}(T)$. This establishes the navigation step in the proof sketch of Section 3.6. Under the hypotheses of Proposition 8, we prove on $\mathcal{C}_T^{\mathrm{ker}}(\zeta) \cap \mathcal{C}_T^{\mathrm{nav}}(\zeta)$ that

$$F_p^{1:J}(\hat{w}(T)) \ge \frac{1}{T_{2\mu}^\star(p)} - r_{\mathrm{nav}}(T,\mu,\zeta), \qquad \lim_{\zeta\downarrow 0} \limsup_{T\to\infty} r_{\mathrm{nav}}(T,\mu,\zeta) = 0.$$

The events are recalled below, and Proposition 11 gives the explicit remainder. This is the objective bound used in Appendix K to control the stopping time.

The proof has two steps. Lemmas 18 and 19 control the changing sampling kernels; Proposition 10 then shows that empirical frequencies track their stationary occupancies with exponentially small failure probability. Next, Lemma 20 converts the flow bound from Appendix H into occupancy tracking, and Lemma 21 supplies the interiority needed for continuity of the information objective. The final proposition combines these facts with the exchange objective bound. The first step uses a Poisson-equation martingale decomposition, as in the non-homogeneous Markov-chain framework of Fort et al. (2011).

Assume Assumption 2 and fix $0 < \mu \le 1/(|\mathcal{S}||\mathcal{A}|)$. Continue the sampling rule after stopping for analysis. With the filtration of Appendix G, the policy for $Z_t \to Z_{t+1}$ and its predictable transition matrix are

$$\chi_t(a \mid s) := \frac{\bar{x}_t(s,a)}{\sum_b \bar{x}_t(s,b)}, \qquad P_t\big((s,a),(s',a')\big) := p(s' \mid s,a)\chi_t(a' \mid s').$$

Thus χ_t is fixed before observing S_{t+1}; the subsequent empirical and exchange updates determine χ_{t+1}. Write $d := |\mathcal{S}||\mathcal{A}|$, $\alpha_t := \alpha_p^{\chi_t}$, and, for $T \geq 1$,

$$\hat{w}_{sa}(T) := N_{sa}(T)/T, \qquad \vartheta_T := \lceil T^{3/4} \rceil, \qquad \widetilde{\alpha}_T^{\vartheta} := \frac{1}{T - \vartheta_T + 1} \sum_{t=\vartheta_T}^{T} \alpha_t,$$

$$\mathcal{C}_T^{\mathrm{nav}}(\zeta) := \{\|\hat{w}(T) - \widetilde{\alpha}_T^{\vartheta}\|_1 \leq \zeta\}.$$

For matrices, use $\|M\|_{1\to 1} := \max_z \sum_{z'} |M(z, z')|$, the induced norm for row vectors. The kernel event $\mathcal{C}_T^{\mathrm{ker}}(\zeta)$ and $\tau_T = \lceil T^{1/4} \rceil$ remain as in Appendix G.

I.1 Slow variation and uniform Poisson bounds

Lemma 18 (Slow variation of the sampling kernels)**.** *Let $A_t := \sum_{u<t} \eta_u$. For $t \geq 1$,*

$$\|\bar{x}_{t+1} - \bar{x}_t\|_1 \leq \frac{2\eta_t}{A_{t+1}}, \qquad \|P_{t+1} - P_t\|_{1\to 1} \leq L_{P,\mu} \frac{\eta_t}{A_{t+1}} \leq \frac{L_{P,\mu}}{t+1},$$

where $L_{P,\mu} = 4/(|\mathcal{A}|\mu)$ is sufficient.

Proof. The averaging identity gives $\bar{x}_{t+1} - \bar{x}_t = (\eta_t/A_{t+1})(x_{t+1} - \bar{x}_t)$, proving the first bound. For $x, y \in K_\mu$, put $x_s = \sum_a x_{sa}$ and $\chi_x(a \mid s) = x_{sa}/x_s$. Since $x_s \geq |\mathcal{A}|\mu$,

$$\sum_a |\chi_x(a \mid s) - \chi_y(a \mid s)| \leq \frac{\sum_a |x_{sa} - y_{sa}| + |x_s - y_s|}{x_s} \leq \frac{2\|x - y\|_1}{|\mathcal{A}|\mu},$$

$$\|P_x - P_y\|_{1\to 1} \leq \max_s \sum_a |\chi_x(a \mid s) - \chi_y(a \mid s)|.$$

Apply these bounds to consecutive averages. Finally, the stepsizes are positive and nonincreasing, so $A_{t+1} \geq (t+1)\eta_t$. □

Lemma 19 (Uniform Poisson bounds)**.** *Let $\mathfrak{P}$ be a family of stochastic matrices on a finite set Z, with common minorization*

$$P^m(z, \cdot) \geq \varepsilon \nu(\cdot) \qquad (P \in \mathfrak{P},\ z \in \mathsf{Z}),$$

where $m \geq 1$, $0 < \varepsilon < 1$, and ν is a probability law. Each P has a unique invariant law α_P. There are constants $H, L < \infty$, depending only on (m, ε), such that for every $\|f\|_\infty \leq 1$ one can choose $g_{P,f}$ with

$$g_{P,f} - P g_{P,f} = f - \alpha_P(f)\mathbf{1},$$
$$\|g_{P,f}\|_\infty \leq H, \qquad \|g_{P,f} - g_{Q,f}\|_\infty \leq L\|P - Q\|_{1\to 1}.$$

Proof. Minorization contracts zero-mass row vectors by $1-\varepsilon$ every m steps. Existence of an invariant law for a finite stochastic matrix therefore gives uniqueness and, with $W_P := \mathbf{1}\alpha_P$,

$$\|P^k - W_P\|_{1\to 1} \leq 2(1 - \varepsilon)^{\lfloor k/m \rfloor}.$$

Consequently,

$$D_P := \sum_{k\geq 0} (P^k - W_P), \qquad \|D_P\|_{1\to 1} \leq H := 2m/\varepsilon, \qquad (I - P)D_P = D_P(I - P) = I - W_P.$$

Set $g_{P,f} = D_P f$. This gives the Poisson equation and its uniform bound. For the Lipschitz bound, stationarity yields

$$\alpha_P - \alpha_Q = \alpha_P(P - Q)D_Q, \qquad \|W_P - W_Q\|_{1\to 1} \leq H\|P - Q\|_{1\to 1}.$$

Since $W_P D_P = D_P W_P = 0$, $R_P := (I - P + W_P)^{-1} = W_P + D_P$ has norm at most $1 + H$. The inverse identity gives

$$D_P - D_Q = R_P[(P - Q) - (W_P - W_Q)]R_Q - (W_P - W_Q).$$

Its norm is at most $L\|P - Q\|_{1\to 1}$, with $L = (1 + H)^3 + H$. Applying it to f completes the proof. □

I.2 Tracking the changing stationary occupancies

Proposition 10 (Quantitative navigation probability bound)**.** *For every fixed $\zeta > 0$, there are $B_{\mathrm{nav}}(\mu, p, \zeta), C_{\mathrm{nav}}(\mu, p, \zeta) > 0$ such that, for all sufficiently large T,*

$$\mathbb{P}((\mathcal{C}_T^{\mathrm{nav}}(\zeta))^c) \le B_{\mathrm{nav}}(\mu, p, \zeta) e^{-C_{\mathrm{nav}}(\mu,p,\zeta)T}.$$

Proof. Fix a coordinate z, put $f = \mathbf{1}_{\{z\}}$, $m = \vartheta_T$, and $n = T - m$. Lemma 8 and Lemma 19 give predictable $h_t = g_{P_t,f}$ with

$$h_t - P_t h_t = f - \alpha_t(f), \qquad \|h_t\|_\infty \le H, \qquad \|h_t - h_{t-1}\|_\infty \le L\|P_t - P_{t-1}\|_{1\to 1},$$

where H, L depend only on (μ, p). In particular, $D_{t+1} := h_t(Z_{t+1}) - P_t h_t(Z_t)$ is a martingale difference with $|D_{t+1}| \le 2H$. Summing the Poisson equation gives

$$\sum_{t=m}^{T-1}(f(Z_t) - \alpha_t(f)) = M_T + R_T, \qquad M_T := \sum_{t=m}^{T-1} D_{t+1},$$

$$R_T = h_m(Z_m) - h_{T-1}(Z_T) + \sum_{t=m+1}^{T-1}(h_t - h_{t-1})(Z_t).$$

Lemma 18 implies

$$|R_T| \le 2H + LL_{P,\mu} \sum_{t=m+1}^{T-1} t^{-1} \le C\log(eT).$$

Discarding the first m counted states changes a coordinate's empirical average by at most m/T; appending α_T to the stationary average changes it by at most $1/(n+1)$. Hence

$$|\hat{w}_z(T) - \widetilde{\alpha}_T^{\vartheta}(z)| \le \frac{|M_T|}{n} + r_T, \qquad r_T := \frac{C\log(eT)}{n} + \frac{m}{T} + \frac{1}{n+1} \longrightarrow 0.$$

For large T, $n \ge T/2$ and $r_T \le \zeta/(2d)$. Azuma–Hoeffding and a union bound over the d coordinates give

$$\mathbb{P}((\mathcal{C}_T^{\mathrm{nav}}(\zeta))^c) \le \sum_z \mathbb{P}\left(|M_T| > \frac{n\zeta}{2d}\right) \le 2d\exp\left(-\frac{\zeta^2 T}{64d^2H^2}\right),$$

which proves the claim. □

I.3 From flow balance to the empirical objective

Lemma 20 (Approximate flow implies occupancy tracking)**.** *There is $C_\mu(p) < \infty$ such that every $x \in K_\mu$ satisfies*

$$\|\alpha_p^{\chi_x} - x\|_1 \le C_\mu(p)\|B_p x\|_1, \qquad (B_p x)_s := \sum_a x_{sa} - \sum_{y,b} x_{yb} p(s \mid y, b).$$

Proof. The action floor $\chi_x \ge \mu$ gives the common minorization of Lemma 8. For every $\|f\|_\infty \le 1$, Lemma 19 supplies $g - P_x g = f - \alpha_p^{\chi_x}(f)\mathbf{1}$ with $\|g\|_\infty \le H$, uniformly in x. Thus

$$|(x - \alpha_p^{\chi_x})f| = |(x - xP_x)g| \le H\|x - xP_x\|_1.$$

Since $(x - xP_x)_{sa} = \chi_x(a \mid s)(B_p x)_s$, its ℓ_1 norm is $\|B_p x\|_1$. Taking the supremum over f proves the claim with $C_\mu(p) = H$. □

Lemma 21 (Uniform interiority of navigation occupancies)**.** *More generally, for every action floor $0 < \mu \le 1/|\mathcal{A}|$, there is $\underline{\alpha}_{\mathrm{nav}}(p, \mu) > 0$ such that every stationary policy $\chi \ge \mu$ has a unique stationary occupancy $\alpha_p^\chi \in K_{\underline{\alpha}_{\mathrm{nav}}(p,\mu)}$.*

Proof. Choose m with $r := \min_{z,z'}(P_p^{\mathrm{Unif}_{\mathcal{A}}})^m(z, z') > 0$. Then $P_p^\chi \ge |\mathcal{A}|\mu P_p^{\mathrm{Unif}_{\mathcal{A}}}$ implies that every entry of $(P_p^\chi)^m$ is at least $(|\mathcal{A}|\mu)^m r$. This gives uniqueness, and stationarity gives the same lower bound on each occupancy coordinate. Take $\underline{\alpha}_{\mathrm{nav}}(p, \mu) = (|\mathcal{A}|\mu)^m r$. □

Proposition 11 (Navigation reduction)**.** *Under the hypotheses of Proposition 8, set* $\underline{\alpha} := \underline{\alpha}_{\mathrm{nav}}(p,\mu)$ *and* $\mu_\star := \frac{1}{2}\min\{\mu, \underline{\alpha}\}$*. Let* $\psi_{p,\mu_\star}$ *be a nondecreasing continuity modulus of* $F_p^{1:J}$ *on* $K_{\mu_\star}$*. For* $0 < \zeta \le \underline{\alpha}/2$*, on* $\mathcal{C}_T^{\mathrm{ker}}(\zeta) \cap \mathcal{C}_T^{\mathrm{nav}}(\zeta)$*, one has* $\hat{w}(T) \in K_{\mu_\star}$ *and, for sufficiently large* T*,*

$$F_p^{1:J}(\hat{w}(T)) \ge \frac{1}{T^\star_{2\mu}(p)} - r_{\mathrm{ex}}^{\mathrm{tail}}(T,\mu,\zeta) - \psi_{p,\mu_\star}\big(\zeta + C_\mu(p) r_{\mathrm{flow}}(T,\zeta)\big).$$

The sum of the two error terms tends to zero in the order $\lim_{\zeta\downarrow 0}\limsup_{T\to\infty}$.

Proof. Work on the stated event and put $\widetilde{x}_T^\vartheta := (T - \vartheta_T + 1)^{-1}\sum_{t=\vartheta_T}^T \bar{x}_t$. Proposition 9 and Lemma 20 give $\|\alpha_t - \bar{x}_t\|_1 \le C_\mu(p) r_{\mathrm{flow}}(T,\zeta)$ for $\vartheta_T \le t \le T$. Averaging and using the navigation event,

$$\|\hat{w}(T) - \widetilde{x}_T^\vartheta\|_1 \le \zeta + C_\mu(p) r_{\mathrm{flow}}(T,\zeta).$$

Lemma 21 gives $\widetilde{\alpha}_T^\vartheta \in K_{\underline{\alpha}}$, so $\hat{w}_{sa}(T) \ge \underline{\alpha} - \zeta \ge \underline{\alpha}/2$. Since $\widetilde{x}_T^\vartheta \in K_\mu$, both allocations lie in $K_{\mu_\star}$. Concavity and Proposition 8 yield

$$F_p^{1:J}(\widetilde{x}_T^\vartheta) \ge \inf_{\vartheta_T \le t \le T} F_p^{1:J}(\bar{x}_t) \ge \frac{1}{T^\star_{2\mu}(p)} - r_{\mathrm{ex}}^{\mathrm{tail}}(T,\mu,\zeta).$$

Apply the continuity modulus to the distance bound. The error limit follows from the two exchange propositions and $\psi_{p,\mu_\star}(r) \to 0$ as $r \downarrow 0$. □

J Continuity of KL minima and positive bounds for stopping

The stopping oracle uses the empirical allocation $\hat{w}(T)$. We first place this allocation in a fixed positive-floor set, then apply the continuity lemma from Appendix H. Compactness converts a fixed gap below the minimum KL value into a uniform positive stopping margin. Throughout, assume Assumption 1 and $p \in \mathcal{P}_{\mathrm{test}}^{1:J}$. The empirical-allocation argument also uses Assumption 2 and the fixed sampling floor $\mu > 0$.

J.1 Positive empirical allocations and local continuity

Recall from Appendix I that

$$\widetilde{\alpha}_T^\vartheta := \frac{1}{T - \vartheta_T + 1}\sum_{t=\vartheta_T}^T \alpha_p^{\chi_t}, \qquad \vartheta_T := \lceil T^{3/4} \rceil, \qquad \mathcal{C}_T^{\mathrm{nav}}(\zeta) := \{\|\hat{w}(T) - \widetilde{\alpha}_T^\vartheta\|_1 \le \zeta\}.$$

The action floor and Lemma 21 give a constant $\underline{\alpha}_{\mu,p} := \underline{\alpha}_{\mathrm{nav}}(p,\mu) > 0$ such that both $\alpha_p^{\chi_t}$ and $\widetilde{\alpha}_T^\vartheta$ have every coordinate at least $\underline{\alpha}_{\mu,p}$. Set $\mu_{\mathrm{nav}} := \underline{\alpha}_{\mu,p}/2$.

Corollary 2 (Interiority of the empirical allocation)**.** *For* $0 < \zeta \le \mu_{\mathrm{nav}}$*, on* $\mathcal{C}_T^{\mathrm{nav}}(\zeta)$*,* $\hat{w}(T) \in K_{\mu_{\mathrm{nav}}}$.

Proof. For every (s,a), $\hat{w}_{sa}(T) \ge \widetilde{\alpha}_T^\vartheta(s,a) - \zeta \ge \underline{\alpha}_{\mu,p} - \zeta \ge \mu_{\mathrm{nav}}$. □

Lemma 22 (Local sign stability)**.** *For every* $p \in \mathcal{P}_{\mathrm{test}}^{1:J}$*, there are* $a_p, \zeta_p > 0$ *such that* $\|p' - p\|_\infty \le \zeta_p$ *implies, for all* $j \in [J]$*,*

$$|V_{p',j}(\rho)| \ge a_p, \qquad \mathrm{sign}(V_{p',j}(\rho)) = \mathrm{sign}(V_{p,j}(\rho)).$$

Proof. Take $a_p = \frac{1}{2}\min_j |V_{p,j}(\rho)| > 0$. Value continuity and finiteness of J give a common neighborhood on which $|V_{p',j}(\rho) - V_{p,j}(\rho)| \le a_p$ for every j. □

For $w \in K_{\mu_{\mathrm{nav}}}$, define

$$g_{p',j}(q) := V_{p',j}(\rho) V_{q,j}(\rho),$$
$$\bar{\sigma}_j(w,p') := \min_{q:\, g_{p',j}(q) \le 0} \Phi(p';w,q),$$
$$u_{\mathrm{NO}}^{(j)}(\sigma,w,p') := \min_{q:\, \Phi(p';w,q) \le \sigma} g_{p',j}(q), \qquad \sigma \ge 0.$$

The last minimum is attained: its feasible set is compact by lower semicontinuity of KL and is nonempty because it contains p'.

Proposition 12 (Local continuity of the minimum weighted KL divergence)**.** *There is $\zeta_{\mathrm{act}} > 0$ such that, for every $j \in [J]$, $\bar{\sigma}_j$ is finite and continuous on $K_{\mu_{\mathrm{nav}}} \times \mathcal{P}_{\zeta_{\mathrm{act}}}(p)$, and on this set*

$$\bar{\sigma}_j(w, p') = F^{(j)}_{p'}(w) = \sigma^{(j)}_{\mathrm{NC}}(0, w, p').$$

Proof. Apply Lemma 15 with allocation floor $\kappa = \mu_{\mathrm{nav}}$, and choose its neighborhood inside that of Lemma 22. The closed alternative sets are then fixed, so $\bar{\sigma}_j = F^{(j)}_{p'}$; Appendix D gives equality with the strict-alternative infimum. The cited continuity proof uses lower semicontinuity of KL and full-support strict competitors, and thus also covers reference kernels with zero transition probabilities. □

J.2 A uniform positive stopping margin

Proposition 13 (Positive stopping objective below the minimum KL value)**.** *Fix $\eta > 0$ and $0 \le \bar{\sigma} < \infty$. There are $\zeta_{\mathrm{sep}}(\eta, \bar{\sigma}) > 0$ and $m_{\mathrm{sep}}(\eta, \bar{\sigma}) > 0$ such that, for every $j \in [J]$, $\|p' - p\|_\infty \le \zeta_{\mathrm{sep}}$, $w \in K_{\mu_{\mathrm{nav}}}$, and $\sigma \in [0, \bar{\sigma}]$,*

$$\sigma \le F^{(j)}_{p'}(w) - \eta \quad \Longrightarrow \quad u^{(j)}_{\mathrm{NO}}(\sigma, w, p') \ge m_{\mathrm{sep}}.$$

Proof. Take $\zeta_{\mathrm{sep}} = \zeta_{\mathrm{act}}$. For each j, consider

$$\mathcal{E}_j := \left\{ (p', w, \sigma, q) \in \mathcal{P}_{\zeta_{\mathrm{act}}}(p) \times K_{\mu_{\mathrm{nav}}} \times [0, \bar{\sigma}] \times \mathcal{P} : \Phi(p'; w, q) \le \sigma \le F^{(j)}_{p'}(w) - \eta \right\}.$$

This set is compact: the KL constraint is closed by lower semicontinuity, and the other constraint is closed by Proposition 12. At every point of $\mathcal{E}_j$, one has $g_{p',j}(q) > 0$. Otherwise q would be a closed alternative and

$$F^{(j)}_{p'}(w) \le \Phi(p'; w, q) \le \sigma \le F^{(j)}_{p'}(w) - \eta,$$

a contradiction. The continuous function $g_{p',j}(q)$ therefore has a strictly positive minimum on each nonempty $\mathcal{E}_j$. Take m_{sep} to be the smallest of these finitely many minima; if all sets are empty, take any positive value. Every feasible kernel in the stated NO problem belongs to the corresponding $\mathcal{E}_j$, so minimizing over that kernel proves the claimed bound. □

K Proof of Theorem 2

We first prove termination and correctness for every positive algorithmic floor. When $\Sigma_p \cap K_{2\mu} \neq \varnothing$, the navigation objective bound and the stopping margin then give a deterministic stopping implication. The concentration bounds make its failure probability summable, yielding the expected-time bound. Finally, we make this argument uniform on compact classes to construct the floor schedule in Corollary 1.

Throughout, assume the hypotheses and implementation requirements of Theorem 2, and fix $p \in \mathcal{P}^{1:J}_{\mathrm{adm}}$. Continue sampling and updating after stopping for analysis, and write

$$\mathcal{C}_T(\zeta) := \mathcal{C}^{\mathrm{ker}}_T(\zeta) \cap \mathcal{C}^{\mathrm{nav}}_T(\zeta), \qquad \Phi(p'; w, q) := \sum_{s,a} w_{sa} \mathrm{KL}_{sa}(p', q).$$

K.1 Termination and the δ-probably correct guarantee

Fix $\delta \in (0, 1)$ and any $0 < \mu \le 1/(|\mathcal{S}||\mathcal{A}|)$. The comparison-set condition $\Sigma_p \cap K_{2\mu} \neq \varnothing$ is not needed here. Recall the stopping budget

$$\beta(t, \delta) := \log(1/\delta) + (|\mathcal{S}| - 1) \sum_{s,a} \log\left(e \left[1 + \frac{N_{sa}(t)}{|\mathcal{S}| - 1} \right] \right),$$
$$\sigma_t := \beta(t, \delta)/(t \vee 1), \qquad \mathcal{Q}_t := \{ q \in \mathcal{P} : \Phi(\hat{p}_t; \hat{w}(t), q) \le \sigma_t \}.$$

Lemma 23 (Termination for every fixed positive floor)**.** *If $\xi^{\mathrm{stop}}_t \to 0$, then $\mathbb{P}_p(\tau_\delta < \infty) = 1$.*

Proof. The persistent action floor holds even when an empirical occupancy domain is empty. The summable bounds in Lemmas 9 and 10, followed by Borel–Cantelli, give almost surely

$$\min_{s,a} \hat{w}_{sa}(t) \geq b > 0 \quad \text{eventually}, \qquad \|\hat{p}_t - p\|_\infty \longrightarrow 0.$$

For the second assertion, apply the kernel bound to a countable sequence of radii tending to zero. Since $\sigma_t \to 0$ at fixed δ, nonnegativity of KL and Pinsker's inequality imply, eventually,

$$\sup_{q \in \mathcal{Q}_t} \|q - \hat{p}_t\|_\infty \leq \sqrt{2\sigma_t / b} \longrightarrow 0.$$

Thus all kernels in $\mathcal{Q}_t$, as well as $\hat{p}_t$, approach p. Continuity and the finitely many nonzero target values at p give $a > 0$ such that, eventually, their policy values have the correct signs and magnitude at least a. Consequently, for every j,

$$u_{\mathrm{NO}}^{(j)}(\sigma_t; \hat{w}(t), \hat{p}_t) \geq a^2, \qquad L_{t,j}^{\mathrm{stop}} \geq a^2 - \xi_t^{\mathrm{stop}} \geq a^2/2 > 0.$$

Coverage and the nonzero-value guards hold by the same argument. Hence the original process stops in finite time almost surely. □

For correctness, apply the transition-KL deviation bound of Al Marjani et al. (2021, Lemma 15, Appendix E.2) to the raw empirical kernel and visit counts of our adaptive trajectory. Relative to the history up to the choice of A_t, the next state has conditional law $p(\cdot \mid S_t, A_t)$, as required by its mixture-martingale proof. With $N_{sa}(t)\mathrm{KL}_{sa}(\hat{p}_t, p) := 0$ when $N_{sa}(t) = 0$, it gives

$$\mathbb{P}_p\left(\exists t \geq 1 : \sum_{s,a} N_{sa}(t)\mathrm{KL}_{sa}(\hat{p}_t, p) > \beta(t, \delta)\right) \leq \delta. \tag{15}$$

Let $\mathcal{E}_\delta$ be the complementary event, on which $p \in \mathcal{Q}_t$ for every $t \geq 1$. At the almost-surely finite stopping time, the certified brackets of Proposition 4 give

$$0 < L_{\tau_\delta, j}^{\mathrm{stop}} \leq u_{\mathrm{NO}}^{(j)}(\sigma_{\tau_\delta}; \hat{w}(\tau_\delta), \hat{p}_{\tau_\delta}) \leq V_{\hat{p}_{\tau_\delta}, j}(\rho) V_{p,j}(\rho) \qquad (j \in [J])$$

on $\mathcal{E}_\delta$. Every returned sign is therefore correct there, and the probability of any error is at most $\mathbb{P}_p(\mathcal{E}_\delta^c) \leq \delta$. Together with termination, this proves Definition 1 for every fixed positive algorithmic floor.

K.2 Deterministic stopping on the good event

For the rate analysis, also assume $\Sigma_p \cap K_{2\mu} \neq \varnothing$, and abbreviate

$$\Lambda := T_{2\mu}^\star(p), \qquad a := \Lambda^{-1} = \max_{w \in \Sigma_p \cap K_{2\mu}} F_p^{1:J}(w) > 0.$$

Let $\underline{\alpha} = \underline{\alpha}_{\mathrm{nav}}(p, \mu)$, $\mu_{\mathrm{nav}} = \underline{\alpha}/2$, and $\mu_\star = \frac{1}{2}\min\{\mu, \underline{\alpha}\}$. Proposition 11 supplies the remainder

$$r_{\mathrm{stop}}(T, \mu, \zeta) := r_{\mathrm{ex}}^{\mathrm{tail}}(T, \mu, \zeta) + \psi_{p,\mu_\star}\big(\zeta + C_\mu(p) r_{\mathrm{flow}}(T, \zeta)\big), \qquad \lim_{\zeta \downarrow 0} \limsup_{T \to \infty} r_{\mathrm{stop}}(T, \mu, \zeta) = 0.$$

For $0 < \zeta \leq \underline{\alpha}/2$, on $\mathcal{C}_T(\zeta)$ and for large T, that proposition and Corollary 2 give

$$\hat{w}(T) \in K_{\mu_\star} \cap K_{\mu_{\mathrm{nav}}}, \qquad F_p^{1:J}(\hat{w}(T)) \geq a - r_{\mathrm{stop}}(T, \mu, \zeta).$$

Lemma 15, applied with floor $\kappa = \mu_\star$, also gives $\zeta_\mu^\circ > 0$ and a modulus $\omega_\mu(\zeta) \to 0$ such that

$$\sup_{\substack{w \in K_{\mu_\star} \\ \|p' - p\|_\infty \leq \zeta}} |F_{p'}^{1:J}(w) - F_p^{1:J}(w)| \leq \omega_\mu(\zeta) \qquad (0 < \zeta \leq \zeta_\mu^\circ).$$

These two bounds provide the objective gap needed for stopping.

Lemma 24 (Deterministic stopping on the good event)**.** *For every $\lambda > \Lambda$, there are $\zeta_\lambda > 0$, $T_\lambda < \infty$, and $\delta_\lambda \in (0, 1)$ such that, for $T \geq T_\lambda$ and $0 < \delta \leq \delta_\lambda$, on $\mathcal{C}_T(\zeta_\lambda)$,*

$$T \geq \lambda \log(1/\delta) \quad \Longrightarrow \quad \tau_\delta \leq T.$$

Proof. Choose $\eta > 0$ with $1/\lambda + 3\eta < a$, and put $\bar{\sigma} = 1/\lambda + \eta$. For $T \geq \lambda \log(1/\delta)$,

$$\sigma_T \leq \frac{1}{\lambda} + \frac{|\mathcal{S}||\mathcal{A}|(|\mathcal{S}|-1)}{T} \log\left(e\left[1 + \frac{T}{|\mathcal{S}|-1}\right]\right) \leq \frac{1}{\lambda} + \eta = \bar{\sigma} \tag{16}$$

for all sufficiently large T. Apply Proposition 13 with gap η, budget bound $\bar{\sigma}$, and allocation floor μ_{nav}, obtaining a radius $\zeta_{\mathrm{sep}} > 0$ and margin $m > 0$. Choose

$$0 < \zeta_\lambda \leq \min\{\zeta_\mu^\circ, \zeta_{\mathrm{sep}}, \zeta_p, \underline{\alpha}/2\}$$

small enough, and then T_λ large enough, that all preceding bounds hold and

$$r_{\mathrm{stop}}(T, \mu, \zeta_\lambda) + \omega_\mu(\zeta_\lambda) \leq \eta, \qquad \xi_T^{\mathrm{stop}} \leq m/2 \quad (T \geq T_\lambda). \tag{17}$$

Here ζ_p is the sign-stability radius of Lemma 22. These choices are independent of δ; one may take any fixed $\delta_\lambda \in (0,1)$.

On $\mathcal{C}_T(\zeta_\lambda)$, the objective bounds now yield

$$F_{\hat{p}_T}^{1:J}(\hat{w}(T)) \geq a - \eta, \qquad \sigma_T \leq 1/\lambda + \eta < a - 2\eta \leq F_{\hat{p}_T}^{1:J}(\hat{w}(T)) - \eta.$$

The joint objective is the minimum over policies, so the same gap holds below every $F_{\hat{p}_T}^{(j)}(\hat{w}(T))$. Moreover, $\hat{w}(T) \in K_{\mu_{\mathrm{nav}}}$, $\|\hat{p}_T - p\|_\infty \leq \zeta_{\mathrm{sep}}$, and $\sigma_T \in [0, \bar{\sigma}]$. The separation proposition therefore gives

$$u_{\mathrm{NO}}^{(j)}(\sigma_T; \hat{w}(T), \hat{p}_T) \geq m, \qquad L_{T,j}^{\mathrm{stop}} \geq m - \xi_T^{\mathrm{stop}} \geq m/2 > 0 \quad (j \in [J]).$$

Positive empirical allocations ensure coverage, and sign stability ensures nonzero reference values. Thus, unless the algorithm has already stopped, it makes these terminating queries at time T and stops then. □

K.3 Stopping tails and the expectation bound

Lemma 25 (Tail bound for the stopping time)**.** *Fix $\lambda > \Lambda$ and $n_\delta := \lceil \lambda \log(1/\delta) \rceil$. There are $B_\lambda, C_\lambda > 0$ such that, for all sufficiently small δ and every integer $t \geq n_\delta$,*

$$\mathbb{P}_p(\tau_\delta > t) \leq f_\lambda(t) := B_\lambda t e^{-C_\lambda t^{1/4}} + B_\lambda e^{-C_\lambda t}.$$

Proof. Choose δ small enough that $n_\delta \geq T_\lambda$ and all concentration bounds apply. The deterministic stopping lemma gives $\{\tau_\delta > t\} \subset \mathcal{C}_t(\zeta_\lambda)^c$. A union bound, Proposition 7, and Proposition 10 give the result, using $\tau_t = \lceil t^{1/4} \rceil \geq t^{1/4}$. □

Proof of Theorem 2. Termination and correctness were proved in Subsection K.1. For the rate bound, fix $\lambda > \Lambda$. The preceding tail envelope is summable and independent of δ, so the tail-sum identity gives

$$\mathbb{E}_p[\tau_\delta] \leq n_\delta + \sum_{t \geq n_\delta} f_\lambda(t) = n_\delta + o(1) \qquad (\delta \downarrow 0).$$

Divide by $\log(1/\delta)$ and let $\lambda \downarrow \Lambda$ to obtain

$$\limsup_{\delta \downarrow 0} \frac{\mathbb{E}_p[\tau_\delta]}{\log(1/\delta)} \leq T_{2\mu}^\star(p).$$

Lemma 11 gives $T_{2\mu}^\star(p) \to T^\star(p)$ as $\mu \downarrow 0$, proving the two-stage limit. □

K.4 A floor schedule independent of the unknown kernel

A pointwise fixed-floor limsup does not by itself give a common schedule: its confidence threshold may depend on p. We first obtain thresholds uniform on nested compact classes. Throughout this subsection, the problem primitives, initialization rules, numerical routines, and deterministic tolerance schedules are fixed independently of p and δ.

Compact classes. Write $N := |\mathcal{S}||\mathcal{A}|$, $P_p^u := P_p^{\mathrm{Unif}_{\mathcal{A}}}$, and define

$$\mathcal{K}_n := \left\{ p \in \mathcal{P} : \min_{z,z'} (P_p^u)^n(z,z') \geq \frac{1}{n}, \quad \min_j |V_{p,j}(\rho)| \geq \frac{1}{n} \right\}, \qquad n \geq 1. \tag{18}$$

These classes are compact by continuity, and lie in $\mathcal{P}_{\mathrm{adm}}^{1:J}$: a positive matrix power gives irreducibility and aperiodicity, and the value bound excludes zero answers. They are nested because left multiplication by a stochastic matrix cannot decrease the smallest entry. Every admissible kernel has a positive power and finitely many nonzero target values, so $\bigcup_n \mathcal{K}_n = \mathcal{P}_{\mathrm{adm}}^{1:J}$. Stationarity of the uniform-policy occupancy also gives

$$\underline{\alpha}_p \geq 1/n \qquad (p \in \mathcal{K}_n). \tag{19}$$

Lemma 26 (Uniform fixed-floor bound on a compact class)**.** *Fix a nonempty $\mathcal{K}_n$ and $0 < \mu \leq \min\{1/(2N), 1/(4n)\}$. With the same positive deterministic tolerances satisfying* (9) *for every kernel and confidence level, for every $\varepsilon > 0$ there is $d(n,\mu,\varepsilon) \in (0,1)$ such that*

$$\frac{\mathbb{E}_p[\tau_{\delta,\mu}]}{\log(1/\delta)} \leq T_{2\mu}^\star(p) + \varepsilon \quad (p \in \mathcal{K}_n,\ 0 < \delta \leq d(n,\mu,\varepsilon)). \tag{20}$$

Proof. We verify that the fixed-floor proof has common constants and remainders on $\mathcal{K}_n$.

Uniform mixing and approximation bounds. The compact set

$$\mathcal{U}_n := \left\{ p' \in \mathcal{P} : \min_{z,z'} (P_{p'}^u)^n(z,z') \geq \frac{1}{2n}, \quad \min_j |V_{p',j}(\rho)| \geq \frac{1}{2n} \right\}$$

contains a neighborhood of $\mathcal{K}_n$. For every $p' \in \mathcal{U}_n$ and every policy $\chi \geq \mu$,

$$(P_{p'}^\chi)^n(z,z') \geq (|\mathcal{A}|\mu)^n (P_{p'}^u)^n(z,z') \geq b_{n,\mu} := \frac{(|\mathcal{A}|\mu)^n}{2n} > 0.$$

The same one-step domination gives the adaptive conditional minorization by the tower argument in Lemma 8. Thus the visit-count, kernel-concentration, occupancy-perturbation, and Poisson bounds have common constants. Every stationary occupancy is bounded below by $b_{n,\mu}$, including those of the comparator policies χ^μ; a continuous choice of optimal allocation is unnecessary.

Set $\kappa := \frac{1}{2}\min\{\mu, b_{n,\mu}\}$. Lemma 15 gives joint continuity of every policywise objective on $\mathcal{U}_n \times K_\kappa$, hence a common modulus by compactness. The anchor bound $\underline{\alpha}_p \geq 1/n \geq 4\mu$ ensures $\Sigma_p \cap K_{2\mu} \neq \varnothing$ on $\mathcal{K}_n$. Uniform occupancy perturbation keeps the transported 2μ-floor comparators feasible once the kernel error is uniformly small. The global cut bound in Lemma 16, the deterministic weight bounds, and the fixed tolerance schedules then make the remainders in Propositions 8, 9, and 11 vanish uniformly in p, first as $T \to \infty$, then as $\zeta \downarrow 0$.

Uniform stopping and expectation bounds. There are $0 < t_- \leq t_+ < \infty$ with $t_- \leq T^\star(p) \leq t_+$ on $\mathcal{K}_n$. Indeed, the fixed full-support alternatives in Lemma 16 uniformly bound the objective above; positivity and continuity of $F_p^{1:J}(\alpha_p^{\mathrm{Unif}_{\mathcal{A}}})$ on $\mathcal{K}_n$ bound it away from zero at the uniform-policy occupancy. Set

$$\Lambda_p := T_{2\mu}^\star(p), \qquad \lambda_p := \Lambda_p + \varepsilon/2.$$

Lemma 11 and $2\mu/\underline{\alpha}_p \leq 1/2$ give $t_- \leq \Lambda_p \leq 2t_+$. Thus λ_p is uniformly bounded above and away from zero, and $1/\Lambda_p - 1/\lambda_p$ has a common positive lower bound. Use this bound to choose one gap $\eta > 0$ and one finite budget bound for the deterministic stopping proof. The compactness proof of Proposition 13, now on $\mathcal{U}_n \times K_\kappa$, gives a common positive stopping margin: the feasible tuples with a fixed gap below the continuous KL minimum form a compact set, on which every value product is positive. Sign stability and the approximation bounds above likewise hold in a common neighborhood of $\mathcal{K}_n$.

The deterministic stopping and concentration arguments therefore yield $B, C, d_0 > 0$, independent of p, such that, for $0 < \delta \leq d_0$ and every integer $t \geq \lambda_p \log(1/\delta)$,

$$\mathbb{P}_p(\tau_{\delta,\mu} > t) \leq f(t) := Bte^{-Ct^{1/4}} + Be^{-Ct}.$$

With $n_{\delta,p} = \lceil \lambda_p \log(1/\delta) \rceil$, tail summation gives $\mathbb{E}_p[\tau_{\delta,\mu}] \leq n_{\delta,p} + \sum_{t \geq n_{\delta,p}} f(t)$. The summable tail tends to zero uniformly because $n_{\delta,p} \geq t_- \log(1/\delta) \to \infty$. After reducing d_0, the tail and rounding error divided by $\log(1/\delta)$ are at most $\varepsilon/2$, uniformly in p. Together with $\lambda_p = \Lambda_p + \varepsilon/2$, this proves the lemma. □

Proof of Corollary 1. Set $\mu_k := 1/(8Nk^2)$. Then $\Sigma_p \cap K_{2\mu_k} \neq \varnothing$ for $p \in \mathcal{K}_k$, and Lemma 11 together with (19) gives

$$T^\star_{2\mu_k}(p) \le \frac{T^\star(p)}{1 - 2\mu_k/\underline{\alpha}_p} \le \left(1 + \frac{1}{4Nk - 1}\right) T^\star(p) \le (1 + 1/k)T^\star(p) \qquad (p \in \mathcal{K}_k).$$

Apply the uniform lemma with error $1/k$, and choose $1 > \delta_1 > \delta_2 > \cdots \downarrow 0$, independently of p, so that

$$\frac{\mathbb{E}_p[\tau_{\delta,\mu_k}]}{\log(1/\delta)} \le (1 + 1/k)T^\star(p) + 1/k \qquad (p \in \mathcal{K}_k,\ 0 < \delta \le \delta_k).$$

Empty classes impose no constraint. Define

$$\mu(\delta) := \begin{cases} \mu_k, & \delta_{k+1} < \delta \le \delta_k, \quad k \ge 1, \\ \mu_1, & \delta_1 < \delta < 1. \end{cases}$$

This schedule is independent of p, lies in $(0, 1/N]$, and tends to zero. Every fixed admissible p belongs to $\mathcal{K}_k$ for all large k, so the displayed inequality gives the required limsup. For each δ, the chosen strategy remains δ-PC on the entire admissible class. That class contains every full-row-support testing kernel, so Theorem 1 supplies the matching liminf and proves the corollary. □

L Algorithm subroutines and online computation

This appendix expands the two subroutines in Algorithm 1 and derives the empirical value update. The target policies and reward are fixed inputs throughout.

L.1 Empirical objects and initialization

After t transitions, let

$$N_{say}(t) := \sum_{u=0}^{t-1} \mathbf{1}\{(s_u, a_u, s_{u+1}) = (s, a, y)\}, \qquad N_{sa}(t) := \sum_y N_{say}(t).$$

Set $\hat{w}_{sa}(t) = N_{sa}(t)/(t \vee 1)$ and $\hat{p}_t(y \mid s, a) = N_{say}(t)/N_{sa}(t)$ on observed rows, with uniform rows otherwise. Thus $\hat{p}_0 = \bar{p}_0 = \mathrm{Unif}_{\mathcal{S}}$. Initialize $x_0 \in K_\mu$, $\bar{x}_0 = x_0$, $G_0 > 0$, and $A_0 = 0$. The safeguard handles any initial flow infeasibility; both the center and its average remain in K_μ on every round.

L.2 The certification subroutine

Algorithm 3 applies the certified solver from Appendix E. Positive empirical weights ensure coverage. Each eligible round makes one bracket call per policy; if certification fails, the main loop collects another transition.

Algorithm 3 `Certify`$(p', w, \sigma, \xi^{\mathrm{stop}})$

if some $w_{sa} = 0$ or some $V_{p',j}(\rho) = 0$ **then**
  return false
end if
for $j = 1, \ldots, J$ **do**
  $(L_j, U_j) \leftarrow \texttt{CertifiedTestOracle}_j(p', w, \sigma, \xi^{\mathrm{stop}})$
end for
return $\min_{j \in [J]} L_j > 0$

L.3 The exchange subroutine

Write $\mathsf{E} = (x, G, \bar{x}, A)$ for the input exchange state and $\mathsf{E}^+ = (x^+, G^+, \bar{x}^+, A^+)$ for the output. Algorithm 4 uses the current kernel $p' = \bar{p}_t$ for the cut and ascent step, and the next kernel $p^+ = \bar{p}_{t+1}$ for the feasibility correction. The occupancy divergence is $\mathrm{KL}(x\|y) = \sum_{s,a} x_{sa} \log(x_{sa}/y_{sa})$.

Algorithm 4 `ExchangeUpdate`$(p', p^+, \mathsf{E}, \mu, \xi^{\mathrm{cut}})$

Unpack $\mathsf{E} = (x, G, \bar{x}, A)$
$D \leftarrow \Sigma_{p'} \cap K_\mu$, $D^+ \leftarrow \Sigma_{p^+} \cap K_\mu$
if $D = \varnothing$ or $D^+ = \varnothing$ **then**
    $g \leftarrow 0$, $G^+ \leftarrow G$, $\eta \leftarrow 1/\sqrt{G^+}$, $x^+ \leftarrow x$
else
    if $x \notin D$ **then**
        Choose $x^\circ \in D$ and set $x \leftarrow x^\circ$
    end if
    $(j, q, g) \leftarrow$ `CutOracle`$(p', x, \xi^{\mathrm{cut}})$
    $G^+ \leftarrow G + \|g\|_\infty^2$, $\eta \leftarrow 1/\sqrt{G^+}$
    $y^+ \in \arg\max_{z \in D}\{\langle g, z\rangle - \eta^{-1}\mathrm{KL}(z\|x)\}$
    $x^+ \in \arg\min_{z \in D^+} \mathrm{KL}(z\|y^+)$
end if
$A^+ \leftarrow A + \eta$, $\bar{x}^+ \leftarrow (A/A^+)\bar{x} + (\eta/A^+)x^+$
return $\mathsf{E}^+ = (x^+, G^+, \bar{x}^+, A^+)$

The empty-domain branch is the retained update: its cut is zero, the accumulator stays unchanged, and the retained center still receives the weight η. The active branch is exactly the two-step KL update analyzed in Appendix H. Resetting an infeasible center changes the point used for that round's cut and ascent step, without resetting its past average or accumulated weights. After either branch, the main procedure normalizes $\bar{x}^+$ to obtain χ_{t+1}; it does not resample the action a_{t+1} already drawn under χ_t.

Feasibility of $\Sigma_{p'} \cap K_\mu$ is a linear program in $d = |\mathcal{S}||\mathcal{A}|$ variables, with $|\mathcal{S}|$ flow equations, normalization, and d lower bounds. A feasible point or an infeasibility result for the current domain can be cached from the preceding round. With one additional initial feasibility check, only the next domain needs a new LP each round; the cached point can also replace an infeasible center when needed. The two occupancy KL-prox problems are convex. The theorem analyzes their exact solutions. Proposition 14 extends its conclusions to exactly feasible numerical updates with summable certified first-order errors; Appendix N.2 implements this extension for the two-state experiment.

The cut requirement and its derivation from near-minimizing alternative kernels are in Appendix F. If all reference values are nonzero, the reduction makes J calls to the routine for computing alternative kernels, with accuracy $\xi^{\mathrm{cut}}/4$, selects the smallest certified approximate value $\widehat{\sigma}_j$, and forms the corresponding rowwise KL vector. If some reference value is zero, the output $(j, p', 0)$ is already valid. Neither the feasibility LP nor the occupancy projections supplies the required alternative kernel or its objective certificate.

L.4 Rank-one maintenance of raw empirical policy values

Let $n = |\mathcal{S}|$. For each fixed target policy, define its empirical state-transition matrix and discounted resolvent by

$$P_{t,j}(s, y) := \sum_a \pi^{(j)}(a \mid s)\hat{p}_t(y \mid s, a), \qquad R_{t,j} := (I - \gamma P_{t,j})^{-1}.$$

With the fixed reward vector $r^{(j)}$ from Section 2.1, its value vector and initial value are

$$v_{t,j} = R_{t,j} r^{(j)}, \qquad V_{\hat{p}_t,j}(\rho) = \rho^\top v_{t,j}.$$

Only row (s_t, a_t) of the raw empirical kernel changes after the next transition, including when that row is observed for the first time. Regard $\Delta_t := \hat{p}_{t+1}(\cdot \mid s_t, a_t) - \hat{p}_t(\cdot \mid s_t, a_t)$ as a column vector and put $b_{t,j} := \pi^{(j)}(a_t \mid s_t)e_{s_t}$. Then

$$P_{t+1,j} = P_{t,j} + b_{t,j}\Delta_t^\top, \qquad I - \gamma P_{t+1,j} = I - \gamma P_{t,j} - \gamma b_{t,j}\Delta_t^\top.$$

The rank-one inverse identity gives

$$R_{t+1,j} = R_{t,j} + \frac{\gamma R_{t,j} b_{t,j} \Delta_t^\top R_{t,j}}{1 - \gamma \Delta_t^\top R_{t,j} b_{t,j}}. \tag{21}$$

Both empirical transition matrices are stochastic and $\gamma < 1$, so both resolvents exist. By the matrix determinant identity, the denominator in (21) is therefore nonzero. If $\pi^{(j)}(a_t \mid s_t) = 0$, this policy's matrix and value remain unchanged.

Initializing all J resolvents takes $O(Jn^3)$ arithmetic operations using dense inversion. Each subsequent update and multiplication by $r^{(j)}$ takes $O(n^2)$ per policy, hence $O(Jn^2)$ per transition, with $O(Jn^2)$ storage. These are exact-arithmetic maintenance costs, not complexity bounds for the certified stopping solves or computations of near-minimizing alternative kernels. In particular, the changing smoothing weight in $\bar{p}_t$ can modify every row, so this single-row calculation does not give the same update cost for values at the stabilized reference kernel.

M Extension to certified numerical updates

Theorem 2 assumes exact occupancy KL-prox updates. This appendix extends its guarantee to the certified numerical updates used in Appendix N. The two optimization problems concerned are precisely those in (6):

$$\begin{aligned} \text{predictor:} \quad & \min_{y \in D_t} \big\{ \mathrm{KL}(y \| x_t) - \eta_t \langle g_t, y \rangle \big\}, \\ \text{correction:} \quad & \min_{x \in D_{t+1}} \mathrm{KL}(x \| y_{t+1}), \qquad D_t = \widetilde{\Sigma}^{\mu}_{\bar{p}_t}. \end{aligned}$$

The predictor is the KL-regularized ascent step on the current occupancy domain; equivalently, it KL-projects the normalized multiplicative update $x_t \exp(\eta_t g_t)$ onto D_t. The correction KL-projects that predictor onto the next occupancy domain D_{t+1}. Proposition 14 relaxes optimality in these two problems, while retaining exact flow, normalization, and floor feasibility. It also specifies the compatible arithmetic changes needed for cuts, averaging, and stopping. The alternative-kernel oracle is a separate optimization problem and must still provide the stated cut certificate.

For a nonempty convex set $D \subseteq K_\mu$, define

$$\mathcal{E}_D(z; b, h) := \max_{v \in D} \langle \log(z/b) - h, z - v \rangle, \qquad z \in D, \quad b \in K_\mu,$$

where logarithms and division are coordinatewise. This is the variational-inequality gap for $\mathrm{KL}(z\|b) - \langle h, z \rangle$; the constant part of the gradient cancels on the simplex. The gap is nonnegative and vanishes exactly at a minimizer. In the predictor use $(D, b, h) = (D_t, x_t, \eta_t g_t)$; in the correction use $(D, b, h) = (D_{t+1}, y_{t+1}, 0)$. Thus a bound on this gap certifies approximate optimality of each occupancy update, rather than merely a small constraint residual.

Proposition 14 (Certified implementation of the upper bound)**.** *Under the other hypotheses of Theorem 2, its conclusions remain valid with the following implementation choices. For each fixed μ, the constants and error bounds below are deterministic and fixed independently of p and δ.*

1. *Use positive nonincreasing smoothing weights $\theta_t \le C_\theta (t \vee 1)^{-1/4}$, with $\theta_0 \le 1$. On active rounds, use nonnegative coefficients g_t satisfying*

$$F_t(w) \le \langle g_t, w \rangle \quad (w \in \Delta(\mathcal{S} \times \mathcal{A})), \qquad \langle g_t, x_t \rangle \le F_t(x_t) + \xi_t^{\mathrm{cut}}.$$

   Set $G_{t+1} = G_t + \|g_t\|_\infty^2$, with $g_t = 0$ on skipped rounds, and choose nonincreasing stepsizes with $c_\eta G_{t+1}^{-1/2} \le \eta_t \le G_{t+1}^{-1/2}$ for a fixed $c_\eta > 0$.

2. *Return exactly feasible $y_{t+1} \in D_t$ and $x_{t+1} \in D_{t+1}$ such that*

$$\mathcal{E}_{D_t}(y_{t+1}; x_t, \eta_t g_t) \le \varepsilon_t, \qquad \mathcal{E}_{D_{t+1}}(x_{t+1}; y_{t+1}, 0) \le \varepsilon_t, \qquad \sum_{t \ge 0} \varepsilon_t < \infty.$$

   Feasibility resets and empty-domain skips are as before; every active numerical solve must terminate with these certificates.

3. *The average may record $r_{t+1} \in K_\mu$ instead of x_{t+1}, with $\|r_{t+1} - x_{t+1}\|_1 \le b_t \to 0$. Use its exact weighted average $\bar{r}_t = \sum_{u<t} \eta_u r_{u+1} / \sum_{u<t} \eta_u$ for navigation, retaining the same predictable action timing.*

Stopping may use the original NO brackets, or a certified interval $[\ell_t, u_t]$ for

$$M_t := tF^{1:J}_{\hat{p}_t}(\hat{w}(t)) - \beta(t,\delta), \qquad u_t - \ell_t \le a_t, \quad a_t/t \to 0,$$

and stop when $\ell_t > 0$, after coverage and nonzero empirical values. With these choices common across instances, the conclusion of Corollary 1 also remains valid.

Proof. Cuts and schedules. The cut majorization gives the same supergradient inequality, and $\mu\|g_t\|_\infty \le \langle g_t, x_t\rangle$ gives the same global bound. The smoothing modification changes only the constant in $d_T(\zeta)$; monotonicity still bounds its total variation by one. The stepsize bounds retain the weight lower bound up to c_η, and

$$\sum_{u=m}^{t-1} \eta_u^2 \|g_u\|_\infty^2 \le \sum_{u=m}^{t-1} \frac{G_{u+1} - G_u}{G_{u+1}} \le \log(G_t/G_m).$$

Perturbing the exchange inequality (Appendix H). For the predictor, the certified inequality $\langle \log(y_{t+1}/x_t) - \eta_t g_t, y_{t+1} - w\rangle \le \varepsilon_t$, $w \in D_t$, replaces exact optimality in Lemma 17. The KL three-point identity therefore adds ε_t to (13). For the correction and $w \in D_{t+1}$, the same identity gives

$$\mathrm{KL}(w\|y_{t+1}) \ge \mathrm{KL}(w\|x_{t+1}) + \mathrm{KL}(x_{t+1}\|y_{t+1}) - \varepsilon_t.$$

Thus the one-step bound acquires only $2\varepsilon_t$. Its cumulative contribution is at most $2\sum_u \varepsilon_u < \infty$, which vanishes after division by $W_{\tau_T:t} \ge C^{-1}T^{3/8}$.

Averaging and navigation (Appendices H and I). Recording r_{u+1} instead of x_{u+1} changes the tail average by at most $\sup_{u\ge\tau_T} b_u$; the prefix still costs $O(T^{-1/8})$. Continuity therefore preserves the objective bound. Exact feasibility of x_{u+1} and $\|B_p(r_{u+1} - x_{u+1})\|_1 \le 2b_u$ preserve the flow bound. Moreover,

$$\|\bar{r}_{t+1} - \bar{r}_t\|_1 \le \frac{2\eta_t}{\sum_{u\le t}\eta_u} \le \frac{2}{t+1},$$

so Appendix I's slow-variation and tracking arguments apply unchanged. All recorded allocations lie in K_μ, retaining the action floor and concentration bounds.

Stopping and the common floor schedule (Appendix K). For the alternative stopping rule, $\ell_t > 0$ excludes every closed sign-flipped alternative from the common plausible set. The time-uniform confidence event in Appendix K.1 therefore gives δ-correctness. The same coverage and consistency argument makes $F^{1:J}_{\hat{p}_t}(\hat{w}(t))$ bounded away from zero eventually, so M_t grows linearly, while $\beta(t,\delta) = O(\log t)$ at fixed δ and $a_t = o(t)$. This proves termination. On the good event in Appendix K.2, for each $\lambda > T^\star_{2\mu}(p)$, the objective and budget bounds give $M_t \ge ct$ for some $c > 0$, whenever $t \ge \lambda\log(1/\delta)$ is sufficiently large. Then $\ell_t \ge M_t - a_t > 0$. The same tail summation proves the expected-time bound. Finally, the new constants and error sequences are common on the compact classes of Appendix K.4; its uniform argument and diagonal construction therefore give the stated corollary. □

N Experimental Details

The experiment implements the certified numerical variant in Proposition 14. Updates and stopping checks occur after every transition. Exact rational arithmetic enforces feasibility; outward-rounded intervals certify optimization and stopping inequalities. The closed-form alternative below is specific to this two-state model. Code, configuration, and verification records are in the supplementary archive.

N.1 Policy values and KL minimization

Fix an action and write $q_s = q(1 \mid s, a)$. Solving the Bellman system gives

$$V^a_q(\rho) = \frac{\gamma(q_0 + q_1 - 1)}{2(1-\gamma)\{1 - \gamma(q_1 - q_0)\}}. \tag{22}$$

The denominator is positive, so the zero-value boundary is $q_0 + q_1 = 1$. For positive weights w_0, w_1, convexity of Bernoulli KL implies that the infimum over opposite-sign alternatives equals the minimum on this boundary. Differentiating its weighted KL objective gives

$$\begin{aligned} q_0^\star = z := \frac{w_0 p_0 + w_1(1-p_1)}{w_0 + w_1}, \qquad q_1^\star = 1 - z, \\ F_p^a(w) = w_0 \,\mathrm{kl}(p_0, z) + w_1 \,\mathrm{kl}(p_1, 1-z). \end{aligned} \tag{23}$$

Other-action rows equal the reference model and incur zero cost. If a reference value is zero, the joint objective and a valid cut are both zero. All boundary coordinates and empirical probabilities are stored as exact fractions, including the cases $p_s = 0$ or 1.

Arithmetic and cut certificates. Outward-rounded interval arithmetic encloses KL values, using correctly rounded logarithms and adjacent representable values. Exact zero terms are handled algebraically; precision increases until the certificates below hold. The supplementary code specifies the arithmetic implementation.

For a cut, enclose both policywise costs in (23), select the smaller upper bound, and round its rowwise KL upper bounds upward to dyadic coefficients $g \geq 0$. Their exact rational witness is feasible. Thus $F(w) \leq \langle g, w \rangle$ for every allocation w. Accept the cut only when $\langle g, x \rangle - \min_j L_j \leq (t+1)^{-2}$, where L_j are the cost lower bounds and t is the number of transitions collected. This certifies the second cut inequality in Proposition 14. Refining the arithmetic and dyadic grid makes the gap vanish.

Stopping certificates. Both methods use the raw empirical model and counts in (23). After coverage and nonzero empirical values, enclose

$$M_t = \min_a F_{\hat{p}_t}^a(N(t)) - \left\{ \log 10 + \sum_{s,a} \big[1 + \log(1 + N_{sa}(t))\big] \right\}.$$

The term in braces is (4) for this experiment. Refine until the interval width is at most $(t+1)^{-2}$, and stop only when its lower endpoint is positive. There is no additional numerical margin. Empirical signs are determined by rational comparisons.

N.2 Certified allocation updates

Initialize the empirical and reference rows to $(1/2, 1/2)$, the center and average to $(1/4, 1/4, 1/4, 1/4)$, and $G_0 = 1$, $A_0 = 0$. The floor is $\mu = 1/50$. The uniform-action stationary occupancy is $(0.325, 0.325, 0.175, 0.175)$, so $\Sigma_p \cap K_{2\mu} \neq \varnothing$. The next action is drawn before updating the reference model and allocation, preserving Algorithm 1's one-step delay. The predictor uses the preceding reference model and the correction uses the new one.

Exact feasibility and certified optimality. In state-major order put $c = (p_{00}, p_{01}, p_{10} - 1, p_{11} - 1)$, where $p_{sa} = p(1 \mid s, a)$. The domain is the polygon

$$D = \{x : \mathbf{1}^\top x = 1,\ c^\top x = 0,\ x_i \geq \mu\}.$$

Enumerate its vertices exactly by fixing two coordinates at μ and solving the two equalities for the others, for every nonsingular pair. The equality matrix has rank two because the reference kernel has full support. Hence this enumerates all vertices and certifies emptiness as well.

For each KL-prox problem, a floating-point solver proposes a point. Reconstruct rational coordinates and solve the two equality constraints exactly, retaining only points in D. For a reconstructed point z, evaluate the interval upper bound

$$\max_{v \in \mathrm{vert}(D)} \langle \log(z/b) - h, z - v \rangle.$$

The maximum over vertices equals the maximum over D. Accept only if this bound is at most $(t+1)^{-2}$, for both the predictor $(b, h) = (x, \eta g)$ and correction $(b, h) = (y, 0)$.

If the proposal fails, enumerate convex combinations of all vertices with dyadic weights of denominator 2^k, choose a point minimizing a certified objective upper bound, and increase both k and precision until the same gap test passes. The grids are dense; uniform convergence of the objective bounds, strict convexity, and continuity of the gradient on K_μ imply convergence to the minimizer and a vanishing gap. Thus this fallback terminates for every positive tolerance, without a polynomial runtime claim. Only exactly empty domains cause a skipped update, as in the algorithm.

Smoothing, stepsizes, and averaging. The implementation uses positive, nonincreasing rational smoothing weights $\theta_t \le t^{-1/4}$, and the largest dyadic stepsize satisfying $\eta_t^2 G_t \le 1$, so $1/(2\sqrt{G_t}) < \eta_t \le 1/\sqrt{G_t}$. Before averaging, it rounds each corrected center x_t to a rational $r_t \in K_\mu$ with $\|r_t - x_t\|_1 \le (t+1)^{-2}$. The weighted sum and total weight are accumulated exactly; navigation uses their statewise normalization. These choices satisfy the smoothing, stepsize, and averaging conditions of Proposition 14; the projection errors are summable. The supplementary code gives the rounding schedules.

N.3 Sampling and computational resources

The initial state is uniform. Rational transition and action probabilities are sampled by refining a uniform-bit interval until the Bernoulli outcome is determined, avoiding floating-point conversion of these probabilities. The probabilistic model assumes independent random bits; the experiment uses a pseudorandom generator. All 100 runs per method stopped before the 200,000-transition cap; none was excluded. Table 1 reports unweighted run averages and stopping-time standard errors.

Runs were serial on an Apple M4 Mac mini with 16 GB RAM, using Python 3.12.14 and NumPy 2.3.5 (Harris et al., 2020) (BSD 3-Clause license). Timings include simulation, certified updates, and in-memory logging, but exclude file writing. The 200 reported runs took 930 seconds in total; pilot and validation computations are documented in the supplement.

N.4 Policywise evidence

Figure 2 plots policywise information divided by $\beta(t, 0.1)$. The first policy is the bottleneck; adaptive allocation raises its evidence faster. For display, evidence is zero before coverage.

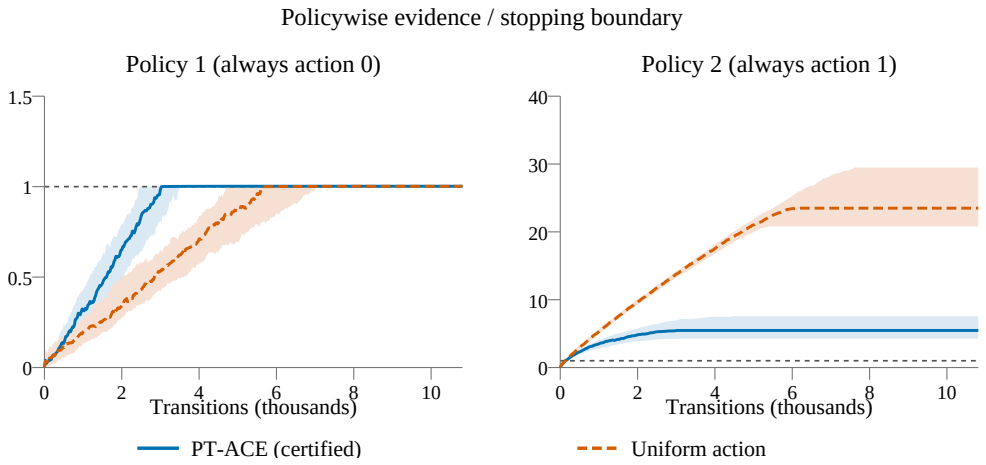


Figure 2: Policywise evidence divided by the stopping boundary. Lines are medians and shading gives interquartile ranges across all 100 runs per method, not confidence intervals. Stopped runs retain their terminal values; no further observations are generated. The horizontal line is one; panels use different vertical scales. Joint stopping requires certified evidence above one for both policies, coverage, and nonzero empirical values.

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: The abstract and introduction state the lower bound, the exchange–navigation contribution, the fixed-floor upper bound, and the existence of a common floor schedule

attaining the exact first-order rate. They explicitly retain nontriviality, anchor-policy ergodicity, certified oracles for stopping and computing near-minimizing alternative kernels, and exact occupancy updates. The two-state experiment implements the certified numerical extension in Proposition 14, with no general empirical or runtime guarantee.

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: Section 5 discusses the instance-dependent floor feasibility condition, the existential confidence thresholds for the common schedule, asymptotic scope, and ergodicity assumption. Section 3.5 states the conditional oracle and exact-update requirements and the absence of a polynomial runtime guarantee. Proposition 14 states the additional requirements for numerical updates. Section 4 limits the empirical comparison to one small instance.

Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: The main assumptions are stated in the main text and referenced in the theorem statements. Formal proofs of the lower bound, exchange analysis, kernel concentration, tracking of target occupancies by empirical frequencies, stopping sensitivity, and the final upper bound are provided in the appendices, with the main text giving a roadmap of how the components fit together.

   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: Appendix N specifies the closed-form KL minimization values, per-transition updates, exact feasibility, certified optimality and stopping tolerances, initialization, sampling, and aggregation. The supplementary archive contains the code, configuration, and validation records; no reported trajectory was excluded.

   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: [No]

   Justification: Appendix N gives the experimental protocol. The implementation, audit scripts, and per-run records are packaged in the supplementary archive, but have not yet been publicly released.

   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: Section 4 gives the MDP and comparison setup. Appendix N describes the certified numerical updates, stopping checks, tolerances, and transition cap; the supplementary code gives the remaining implementation settings.

   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: [Yes]

   Justification: Table 1 reports stopping-time means and standard errors across 100 independent trajectories per method. The standard error is the sample standard deviation (denominator 99) divided by $\sqrt{100}$. We report observed joint errors and make no formal significance or general empirical-superiority claim.

   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: Appendix N.3 reports the CPU, RAM, Python and NumPy versions, serial execution, timing scope, and aggregate runtime of the reported runs. Additional pilot and validation computations are documented in the supplement.

   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 work consists of theory and a small synthetic experiment; it does not involve human subjects, personal data, sensitive datasets, or deployed systems. To the best of our knowledge, the research and its presentation conform to the NeurIPS Code of Ethics.

   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: [Yes]

    Justification: The conclusion includes a broader-impact discussion. It explains that the paper provides methodological tools for statistically reliable policy testing, which may support evidence-based decision making in applications such as recommendation, operations, and treatment-strategy design. It also notes that the guarantees certify only the value criteria specified by the reward, target policies, and model class, and therefore do not by themselves ensure social desirability, fairness, or robustness to unmodeled shifts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: The paper does not release datasets, pre-trained models, generative models, or other artifacts that pose high misuse risks.

    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: [Yes]

    Justification: The synthetic experiment uses NumPy 2.3.5, credited in Appendix N.3 and used under its BSD 3-Clause license (`https://numpy.org/doc/stable/license.html`). No external dataset or pretrained model is used.

    Guidelines:

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

13. **New assets**

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

    Answer: [N/A]

    Justification: The paper does not release new research assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, human-subject experiments, or user studies.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: The paper does not involve human subjects, so IRB approval or equivalent review is not applicable.

    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: No LLM is a component of the proposed testing method. LLM assistance was used for manuscript editing and implementation of the synthetic experiment. The supplementary archive documents the analytical checks and trajectory-based audit of that implementation.

    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.