Adaptive Joint Testing of Policies in Discounted Markov Decision Processes

Po-An Wang [1] **Kaito Ariu** [2 3]

Abstract

We study fixed-confidence joint policy testing in discounted tabular Markov decision processes under active exploration. Given a finite family of target policies, the learner observes a single adaptive trajectory and must certify the sign of every target-policy value with probability at least $1-\delta$. We identify the instance-specific first-order benchmark for this problem: a characteristic time $T^\star(p)$, defined by a max–min program over stationary occupancy measures and sign-flipped alternatives. We then develop PT-ACE(μ), an online algorithm that couples stabilized exchange-based learning of a bottleneck occupancy allocation, trajectory-compatible navigation from averaged occupancies, and parallel certified policywise stopping via scalar frontier tests. Under a trajectory-side anchor-policy ergodicity condition, and when run with certified numerical subroutines and vanishing input tolerances, PT-ACE(μ) is δ-correct and, for every sufficiently small admissible fixed floor $\mu > 0$, satisfies

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

where $c(\mu, p) \to 0 \quad \text{as } \mu \downarrow 0$. Thus a single adaptive trajectory can certify multiple policy signs at the instance-specific lower-bound rate in the vanishing-floor limit.

1. Introduction

Policy testing asks whether the value of a target policy has a prescribed sign under an unknown environment. In many applications, however, one does not test only a single candidate policy. A practitioner may have a finite collection of controllers, treatment rules, or recommendation policies, and may wish to certify, from one stream of interaction data, which of them have positive value.

This paper studies joint zero-threshold policy testing in discounted tabular Markov decision processes. Let $\Pi_{\mathrm{test}} := \{\pi^{(1)}, \dots, \pi^{(J)}\}$ be a fixed finite family of target policies. Under a common known reward function, the goal is to determine the sign of each value $V_{p,j}(\rho)$, $j \in [J]$, under the unknown transition kernel p. Equivalently, the learner must output $\mathrm{Ans}_{1:J}(p) := (\mathrm{Ans}_1(p), \dots, \mathrm{Ans}_J(p))$, where $\mathrm{Ans}_j(p) := \mathrm{sign}(V_{p,j}(\rho))$, with probability at least $1-\delta$.[1]

This problem is both a statistical inference problem and an online allocation problem. Some state-action pairs may be highly informative for one policy but nearly irrelevant for another. Since all tests share the same trajectory, the learner must collect data according to a single exploration strategy that supports the most difficult active test in the family. The optimal first-order sample complexity is therefore instance-specific: each transition kernel p induces a characteristic time $T^\star(p)$, obtained by maximizing a bottleneck testing frontier over stationary occupancy measures.

Appendix A gives a reset-chain example showing that on-policy testing can require exponentially many samples, while active navigation can probe the informative transition row in polynomial time. This motivates the active-allocation viewpoint adopted in the main algorithm.

The online nature of the problem creates a gap between this lower-bound benchmark and an implementable algorithm. The lower bound is expressed in terms of an optimal occupancy allocation, but the learner does not directly choose state-action counts; it chooses actions along one evolving trajectory. Moreover, the feasible occupancy polytope depends on the unknown transition kernel. Thus an asymptotically optimal algorithm must identify where information should be collected, realize that allocation through the controlled dynamics, and maintain valid stopping certificates.

We address these challenges with PT-ACE(μ), short for *Policy Testing via AdaGrad and Cutting-plane Exploration*. The algorithm has three coupled layers. The exchange layer learns an informative allocation for the shared bottleneck testing objective, protecting the hardest policywise sign test in the family. The navigation layer converts averaged

[1]National Tsing Hua University [2]CyberAgent [3]Tohoku University. Correspondence to: Po-An Wang <po-an@stat.nthu.edu.tw>.

ICML 2026 Workshop on Hypothesis Testing, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

[1]Nonzero thresholds can be handled by the standard reward-shifting reduction. For policy-specific thresholds this creates policy-specific shifted rewards; we state the main theory for a common reward and zero thresholds to keep the notation focused.

exchange iterates into an executable behavior policy, so that the single trajectory tracks the learned occupancy target. The stopping layer invokes the certified fixed-policy testing solver of Ariu et al. (2025), separately for each target policy, and stops only when all policywise lower certificates are positive.

The central technical challenge lies in the exchange–navigation layer. The exchange updates are carried out over empirical occupancy domains that move with the data and impose a positive floor on every state-action coordinate. Such domains may be empty during an initial transient, and their geometry changes as the empirical kernel is updated. The bottleneck cuts are also delicate: a sign-flipping alternative need not have bounded rowwise KL merely because it flips a policy value. We handle these issues through a skip–reanchor safeguard, a uniform cut-bound argument based on full-support opposite-sign comparators, and transported interior benchmarks that remain feasible for the moving domains on the tail event.

Contributions. First, we identify the information-theoretic benchmark for online joint policy testing. We define an instance-specific characteristic time $T^\star(p)$ by a max–min program over stationary occupancy measures and sign-flipped alternatives, and prove the corresponding asymptotic lower bound: every δ-probably correct strategy satisfies

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

Second, we develop PT-ACE(μ), an online active-testing algorithm designed to approach this benchmark from a single trajectory. The algorithm combines three ingredients: an exchange layer that learns a bottleneck sampling allocation, a navigation layer that turns this allocation into an executable behavior policy, and a certified stopping layer that tests the target policies in parallel.

Third, we prove that the exchange layer learns a near-optimal bottleneck allocation despite operating over empirical domains that move with the data and impose a positive floor on every state-action coordinate. The analysis shows that the cut vectors remain uniformly controlled, and that the averaged exchange iterates achieve the optimal joint testing frontier up to a vanishing error and a floor-dependent factor.

Fourth, we prove that the learned allocation is realized by the online trajectory. The navigation analysis shows that the empirical allocation generated by the single adaptive trajectory inherits the same bottleneck frontier value as the exchange averages, up to vanishing error.

Finally, combining exchange, navigation, and imported policywise certified stopping gives, for every sufficiently small admissible floor $\mu > 0$,

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_p[\tau_\delta]}{\log(1/\delta)} \leq (1 + c(\mu, p))T^\star(p),$$

where $c(\mu, p) \to 0$ as $\mu \to 0$. Thus PT-ACE(μ) is first-order optimal up to an interiorization factor for each admissible fixed floor, and the family $\{\text{PT-ACE}(\mu)\}_{\mu>0}$ recovers the exact instance-specific characteristic time in the vanishing-floor limit.

Scope of the upper bound. The lower bound is stated over the full stationary occupancy set and does not require an interior floor. The online upper bound uses two proof-side regularizations. First, PT-ACE(μ) keeps the learned occupancies away from the boundary by imposing a positive floor on every state-action coordinate. For a given instance p, the proof requires this floor to be sufficiently small so that the corresponding interior benchmarks remain feasible for the moving empirical domains on the tail event. The algorithm can be run with any floor satisfying $|\mathcal{S}||\mathcal{A}|\mu \leq 1$, but the guarantee is proved for floors in an instance-dependent admissible range. Second, the navigation proof assumes that the uniform-action state-action chain is irreducible and aperiodic, which gives a common mixing mechanism for all positive-floor navigation policies. These upper-bound conditions are not needed for the policywise stopping certificates or for the lower-bound benchmark $T^\star(p)$, but they allow us to prove that the learned allocation is realized by a single adaptive trajectory.

1.1. Related work

Policy testing and static sampling. The closest work to ours is the fixed-policy testing framework of Ariu et al. (2025). They study policy testing in discounted tabular MDPs under static sampling, derive an instance-specific lower bound, and introduce the scalar-frontier and reversed-MDP machinery that makes the policywise stopping problem computationally accessible. We build directly on this policywise certification interface. Our setting differs in two ways: we test a finite family of policies jointly, leading to the bottleneck frontier $F_p^{1:J}$, and samples are generated by a single adaptive trajectory rather than drawn from a prescribed static allocation. The main contribution of the present paper is therefore the online exchange–navigation layer that learns and realizes a shared occupancy allocation while retaining certified policywise stopping.

Online pure exploration and navigation. Instance-dependent pure exploration in MDPs is substantially harder online than under a generative model, because the learner controls actions but not state-action counts directly. This navigation difficulty is central in best-policy identification. Al Marjani et al. (2021) introduce navigation-constrained al-

location sets and online algorithms whose sampling frequencies track target occupancies through non-homogeneous Markov chains. Our navigation proof is conceptually aligned with this line of work, but the statistical objective is different: rather than identifying an optimal policy, we certify a vector of policy-testing answers, and the relevant allocation is the maximizer of a joint testing frontier.

Pure exploration and thresholding. The lower-bound-and-allocation viewpoint follows the broader pure-exploration tradition in bandits, including best-arm identification (Garivier and Kaufmann, 2016; Kaufmann et al., 2016) and game-based or exchange-style algorithms (Degenne et al., 2019; Wang et al., 2021). Policy testing is also related to thresholding bandits (Locatelli et al., 2016), where the goal is to identify which arms exceed a threshold. Our problem can be viewed as a stateful analogue in which each coordinate corresponds to the sign of a fixed policy value, but observations are coupled through a shared unknown transition kernel and must be collected along a single controlled trajectory.

Multi-policy and multi-reward objectives. Recent work has studied multi-reward or multi-policy objectives in MDP pure exploration, including multi-reward best-policy identification (Russo and Vannella, 2024) and adaptive exploration for multi-reward multi-policy evaluation (Russo and Pacchiano, 2025). These works highlight that one trajectory or one dataset may need to support several downstream value questions. Our focus is complementary: the policies are fixed, the goal is joint fixed-confidence sign certification, and the instance-specific benchmark is a bottleneck testing frontier over sign-flipped alternatives.

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 $p \in \mathcal{P}$, 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 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 $\hat{\imath} \in \{+,-\}^J$.

Definition 2.1 (δ-probably correct joint testing)**.** A strategy is δ-probably correct, or δ-PC, if for every $p \in \mathcal{P}_{\text{test}}^{1:J}$, $\mathbb{P}_p(\hat{\imath} \neq \text{Ans}_{1:J}(p)) \leq \delta$.

2.1. Nontriviality and confusing alternatives

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

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 the sign vector under q differs from that under p in at least one coordinate.

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:

$$\Sigma_p := \{\omega \in \mathbb{R}_+^{\mathcal{S}\times\mathcal{A}} : (2) \text{ holds}\}, \tag{1}$$

where

$$\sum_{s,a} \omega_{sa} = 1, \qquad \sum_b \omega_{yb} = \sum_{s,a} \omega_{sa} p(y \mid s,a) \quad \forall y \in \mathcal{S}. \tag{2}$$

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 $\omega \in \Sigma_p$ and $q \in \mathcal{P}$, write

$$\Phi(p;\omega,q) := \sum_{s,a} \omega_{sa} \text{KL}_{sa}(p,q),$$

$$\text{KL}_{sa}(p,q) := \text{KL}(p(\cdot \mid s,a) \,\|\, q(\cdot \mid s,a)).$$

This is the information rate at which an allocation ω separates the true kernel p from an alternative kernel q. For policy j, define

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

the least information collected against alternatives that flip the sign of the j-th target policy. Since the joint test is wrong whenever any coordinate is flipped, the relevant frontier is the bottleneck over policies:

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

The characteristic time is

$$\frac{1}{T^\star(p)} := \max_{\omega \in \Sigma_p} F_p^{1:J}(\omega). \tag{3}$$

Equivalently, $T^\star(p)$ is the inverse of the largest worst-case information rate achievable by any stationary occupancy allocation.

2.3. Asymptotic lower bound

Theorem 2.3 (Asymptotic lower bound for joint testing)**.** *For any δ-PC strategy,*

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

The proof is deferred to Appendix B. Theorem 2.3 identifies the two tasks that any asymptotically optimal algorithm must solve: it must learn an allocation attaining the characteristic-time frontier in (3), and it must determine from data when the accumulated evidence rules out every sign-flipped alternative in the family.

3. PT-ACE(μ): online exchange, navigation, and certification

The lower bound identifies the first-order benchmark $1/T^\star(p) = \max_{\omega\in\Sigma_p} F_p^{1:J}(\omega)$. Attaining it online requires learning an informative occupancy allocation for $F_p^{1:J}$, while also deciding from data when every coordinate of $\mathrm{Ans}_{1:J}(p)$ is certified. The first task is online-specific: p is unknown, the feasible occupancy set moves with the data, and the learner controls only a single trajectory. The second task is handled through the certified policywise stopping theorem of Ariu et al. (2025), invoked in parallel across the finite family.

PT-ACE(μ) addresses these difficulties through three coupled components. The exchange component learns where information should be collected for the bottleneck frontier. The navigation component converts the current learned allocation into an executable behavior policy and uses it to generate the next transition. The stopping component periodically applies policywise certified tests and stops only when all target-policy signs are certified.

At a high level, the algorithm repeats the following procedure. It uses the current averaged exchange allocation to define a navigation policy, executes one transition under this policy, updates the empirical transition model, and then uses the updated model to refine the exchange allocation. The same trajectory therefore plays two roles: it supplies the data used to learn where future samples should be collected, and it provides the evidence used by the stopping certificates. The full update is given in Algorithm 1; before stating it, we describe the three design principles behind the algorithm.

3.1. Design principles

Bottleneck exchange-based allocation learning. The joint testing objective $\omega \mapsto F_p^{1:J}(\omega) = \min_{j\in[J]} \inf_{q\in\mathrm{Alt}_j(p)} \Phi(p;\omega,q)$ is a pointwise minimum of linear forms in the occupancy variable. Thus, at a current occupancy surrogate, a cut oracle identifies a bottleneck policy index and a confusing alternative, generating a supporting linear functional for the outer allocation problem. Concretely, the cut oracle approximately minimizes $\Phi(\bar{p}_t; x_t, q)$ over policy indices and closed sign-flipped alternatives, then returns the rowwise KL vector $g_t = (\mathrm{KL}_{sa}(\bar{p}_t, q_t))_{s,a}$. PT-ACE($\mu$) avoids resolving the full max–min program from scratch at each round and instead performs stabilized KL-prox updates driven by successive bottleneck cuts. The cut oracle is instantiated from the policywise certified frontier solver with witness extraction in Appendix E; the exchange proof uses only the certified bottleneck-cut contract stated in Section 3.6, with tolerances satisfying (4).

Moving domains, safeguards, and navigation. The empirical occupancy polytope $\Sigma_{\hat{p}_t}$ changes with the data and may not contain a point with a prescribed positive floor. For a floor parameter $\mu > 0$, define

$$K_\mu := \{w \in \Delta(\mathcal{S}\times\mathcal{A}) : w_{sa} \geq \mu \;\forall (s,a)\}.$$

Given a kernel p', write

$$\widetilde{\Sigma}^{\mu}_{p'} := \Sigma_{p'} \cap K_\mu$$

for the corresponding fixed-floor occupancy domain. PT-ACE(μ) uses the stabilized empirical kernel

$$\bar{p}_t := (1-\theta_t)\hat{p}_t + \theta_t \mathrm{Unif}_{\mathcal{S}}, \qquad \theta_t = (t\vee 1)^{-1/4},$$

and performs exchange updates over $\widetilde{\Sigma}^{\mu}_{\bar{p}_t}$.

If either the current or next fixed-floor domain is empty, the exchange step is skipped and the current center is retained; if the current domain is nonempty but the stored center is infeasible, the center is reanchored before the cut oracle is queried. On the kernel-concentration event, these safeguards are inactive on the tail window; see Appendix G. Since the learner cannot directly impose target state-action frequencies, PT-ACE(μ) converts a weighted average of corrected or retained exchange centers into a behavior policy. This average lies in K_μ, so the executed policy has a persistent action floor. The non-homogeneous Markov-chain tracking argument and frontier transfer are proved in Appendix H.

Certified policywise stopping. For stopping, PT-ACE(μ) uses the certified fixed-policy testing solver of Ariu et al. (2025), invoked separately for each target policy. The joint rule stops once all policywise lower certificates are positive. Thus the new technical burden is not to redesign the static stopping solver, but to learn and realize the right online allocation for the joint bottleneck objective. The policywise stopping interface is stated in Section 3.5.

3.2. From exchange averages to navigation policies

At round t, the algorithm maintains the raw empirical kernel $\hat{p}_t$, the surrogate kernel $\bar{p}_t$, an exchange center $x_t \in K_\mu$,

and a weighted average $\bar{x}_t \in K_\mu$ of corrected or retained centers. We reserve $\pi^{(j)}$, $j \in [J]$, for fixed target policies. The frozen navigation policy χ_t used for $Z_t \to Z_{t+1}$ is induced by $\bar{x}_t$:

$$\chi_t(a \mid s) := \frac{\bar{x}_t(s,a)}{\sum_{b\in\mathcal{A}} \bar{x}_t(s,b)}.$$

We use a one-step delayed convention: χ_t is fixed before the transition from $Z_t = (s_t, a_t)$ to $Z_{t+1} = (s_{t+1}, a_{t+1})$. Thus, after observing s_{t+1}, the next action a_{t+1} is sampled from $\chi_t(\cdot \mid s_{t+1})$. The empirical update based on (s_t, a_t, s_{t+1}) and the exchange correction then define $\bar{x}_{t+1}$, hence χ_{t+1}. Before the stopping interface is activated, PT-ACE(μ) enforces an initial coverage phase: the stopping rule is queried only after every state-action pair has been observed at least once.

If the exchange step is skipped, no bottleneck cut is queried and the current center is retained. We denote this bookkeeping operation by `RetainUpdate`$(x_t, G_t, \bar{x}_t, A_t)$. It sets

$$g_t := 0, \qquad G_{t+1} := G_t,$$
$$\eta_t := \frac{1}{\sqrt{G_{t+1}}}, \qquad x_{t+1} := x_t,$$

and still incorporates the retained center into the running average:

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

Thus skipped rounds do not introduce new cut information, but they preserve the same averaging convention used on active exchange rounds.

3.3. Near-optimality of the exchange layer

The exchange analysis has to overcome two difficulties: the feasible set is the moving empirical domain $\widetilde{\Sigma}^{\mu}_{\bar{p}_t}$, and the cut oracle returns a sign-flipping alternative whose rowwise KL need not be uniformly bounded merely because it flips a sign. The following informal propositions summarize the key ingredients; formal statements are in Appendix G.

Proposition 3.1 (Uniform control of bottleneck cuts, informal)**.** *Suppose the exchange center lies in K_μ and the cut-oracle tolerances are uniformly bounded. Then $g_t(s,a) = \mathrm{KL}_{sa}(\bar{p}_t, q_t)$ satisfies $\|g_t\|_\infty \le L_\mu$ for all t, for a finite L_μ.*

The proof uses fixed full-support opposite-sign comparators, supplied by nontriviality, to bound the frontier value. The interior floor $x_t(s,a) \ge \mu$ converts a bound on $\langle g_t, x_t\rangle$ into a coordinatewise ℓ_∞ bound.

Proposition 3.2 (Tail-window exchange near-optimality, informal)**.** *Let $\tau_T := \lceil T^{1/4}\rceil$ and $\vartheta_T := \lceil T^{3/4}\rceil$. On $\mathcal{C}^{\mathrm{ker}}_T(\zeta)$, for all sufficiently large T, $\inf_{t\in[\vartheta_T,T]} F^{1:J}_p(\bar{x}_t)$ is larger than*

$$(1 + c(\mu,p))^{-1} \max_{w\in\Sigma_p} F^{1:J}_p(w) - r^{\mathrm{tail}}_{\mathrm{ex}}(T,\mu,\zeta),$$

where $\lim_{\zeta\downarrow 0}\limsup_{T\to\infty} r^{\mathrm{tail}}_{\mathrm{ex}}(T,\mu,\zeta) = 0$.

The comparator is not a fixed optimizer of Σ_p, because the domain moves with $\bar{p}_t$. Instead, we construct a fixed interior navigation policy χ^μ from $w^{2\mu} \in \Sigma_p \cap K_{2\mu}$, and compare against $w^\mu_t := \alpha^{\chi^\mu}_{\bar{p}_t} \in \widetilde{\Sigma}^{\mu}_{\bar{p}_t}$ on the tail good event. This transported comparator has logarithmic total drift, while the cumulative AdaGrad weight diverges uniformly on $[\vartheta_T, T]$.

3.4. Realizing the learned allocation

The exchange layer controls the averaged occupancy targets $\bar{x}_t$, but the algorithm is judged by the empirical allocation $\hat{w}(T)$ generated by a single adaptive trajectory. The navigation layer bridges this gap.

Proposition 3.3 (Navigation transfer, informal)**.** *On $\mathcal{C}^{\mathrm{ker}}_T(\zeta) \cap \mathcal{C}^{\mathrm{nav}}_T(\zeta)$, for all sufficiently large T, $F^{1:J}_p(\hat{w}(T))$ is greater than*

$$(1 + c(\mu,p))^{-1} \max_{w\in\Sigma_p} F^{1:J}_p(w) - r_{\mathrm{nav}}(T,\mu,\zeta),$$

where $\lim_{\zeta\downarrow 0}\limsup_{T\to\infty} r_{\mathrm{nav}}(T,\mu,\zeta) = 0$.

The proof is given in Appendix H. It combines a common minorization from the persistent action floor with an approximate-flow argument: the stationary occupancy of the policy induced by $\bar{x}_t$ is close to $\bar{x}_t$ whenever the p-flow defect of $\bar{x}_t$ is small. This transfers the exchange frontier bound from $\bar{x}_t$ to the realized empirical allocation $\hat{w}(T)$.

3.5. Policywise stopping rule

The exchange and navigation layers explain how PT-ACE(μ) learns and realizes an informative empirical allocation. It remains to convert this allocation into a statistically valid joint decision. We do this through policywise certified scalar-frontier tests, invoked in parallel across the finite family.

The policywise scalar-frontier facts are recalled in Appendix C; the deterministic certified bracket is instantiated in Appendix D; and the witness form used by the exchange cut oracle is described in Appendix E. For fixed j, p, ω, σ, define $u^{(j)}_{\mathrm{NO}}(\sigma;\omega,p)$ as

$$\inf\Big\{V_{p,j}(\rho)V_{q,j}(\rho) : q \in \mathcal{P},\ \Phi(p;\omega,q) \le \sigma\Big\}.$$

If this value is positive, every kernel in the weighted-KL plausible set $\{q : \Phi(p;\omega,q) \le \sigma\}$ has the same j-th sign as p.

Algorithm 1 PT-ACE(μ)

Require: confidence δ, floor $\mu > 0$, target policies $\{\pi^{(j)}\}_{j=1}^{J}$, initial $x_0 \in K_\mu$, tolerances $(\xi_t^{\mathrm{cut}})_{t\geq 0}$, $(\xi_t^{\mathrm{stop}})_{t\geq 0}$
Initialize $\hat{p}_0, \bar{p}_0, x_0, \bar{x}_0 = x_0, \hat{w}(0), N_{sa}(0), G_0 > 0, A_0 = 0$.
Set $\chi_0(a \mid s) := \bar{x}_0(s,a)/\sum_b \bar{x}_0(s,b)$.
Sample $s_0 \sim \rho$ and $a_0 \sim \chi_0(\cdot \mid s_0)$.
for $t = 0, 1, 2, \ldots$ **do**
  Set $\sigma_t \leftarrow \beta(t,\delta)/(t \vee 1)$.
  if $N_{sa}(t) \geq 1$ for all (s,a), and $V_{\hat{p}_t,j}(\rho) \neq 0$ for all j **then**
    for $j = 1, \ldots, J$ **do**
      $(L_{t,j}^{\mathrm{stop}}, U_{t,j}^{\mathrm{stop}}) \leftarrow \texttt{CertifiedTestOracle}_j(\hat{p}_t, \hat{w}(t), \sigma_t, \xi_t^{\mathrm{stop}})$.
    end for
    if $\min_{j\in[J]} L_{t,j}^{\mathrm{stop}} > 0$ **then**
      return $\hat{\imath}_t := (\mathrm{Ans}_1(\hat{p}_t), \ldots, \mathrm{Ans}_J(\hat{p}_t))$.
    end if
  end if
  Observe $s_{t+1} \sim p(\cdot \mid s_t, a_t)$.
  Sample $a_{t+1} \sim \chi_t(\cdot \mid s_{t+1})$.
  Update $N_{sa}(t+1)$, $\hat{w}(t+1)$, and $\hat{p}_{t+1}$ using (s_t, a_t, s_{t+1}).
  Set $\bar{p}_{t+1} \leftarrow (1-\theta_{t+1})\hat{p}_{t+1} + \theta_{t+1}\mathrm{Unif}_{\mathcal{S}}$.
  if $\widetilde{\Sigma}_{\bar{p}_t}^{\mu} = \varnothing$ or $\widetilde{\Sigma}_{\bar{p}_{t+1}}^{\mu} = \varnothing$ **then**
    $(x_{t+1}, G_{t+1}, \eta_t, \bar{x}_{t+1}, A_{t+1}) \leftarrow \texttt{RetainUpdate}(x_t, G_t, \bar{x}_t, A_t)$.
    Set $\chi_{t+1}(a \mid s) := \bar{x}_{t+1}(s,a)/\sum_b \bar{x}_{t+1}(s,b)$.
    continue
  end if
  if $x_t \notin \widetilde{\Sigma}_{\bar{p}_t}^{\mu}$ **then**
    Choose $x_t^{\circ} \in \widetilde{\Sigma}_{\bar{p}_t}^{\mu}$ and set $x_t \leftarrow x_t^{\circ}$.
  end if
  $(j_t, q_t, g_t) \leftarrow \texttt{CutOracle}(\bar{p}_t, x_t, \xi_t^{\mathrm{cut}})$.
  $G_{t+1} \leftarrow G_t + \|g_t\|_\infty^2$, $\eta_t \leftarrow 1/\sqrt{G_{t+1}}$.
  $y_{t+1} \in \arg\max_{x\in\widetilde{\Sigma}_{\bar{p}_t}^{\mu}} \{\langle g_t, x\rangle - (1/\eta_t)\mathrm{KL}(x\|x_t)\}$.
  $x_{t+1} \in \arg\min_{x\in\widetilde{\Sigma}_{\bar{p}_{t+1}}^{\mu}} \mathrm{KL}(x\|y_{t+1})$.
  $A_{t+1} \leftarrow A_t + \eta_t$.
  $\bar{x}_{t+1} \leftarrow (A_t/A_{t+1})\bar{x}_t + (\eta_t/A_{t+1})x_{t+1}$.
  Set $\chi_{t+1}(a \mid s) := \bar{x}_{t+1}(s,a)/\sum_b \bar{x}_{t+1}(s,b)$.
end for

Proposition 3.4 (Certified policywise stopping contract)**.** *Fix $j \in [J]$. For every reference kernel $p \in \mathcal{P}$, strictly positive allocation ω, budget $\sigma \geq 0$, and tolerance $\xi^{\mathrm{stop}} > 0$, the policywise certified frontier routine for $\pi^{(j)}$ returns $L_j(\sigma)$ and $U_j(\sigma)$ such that*

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

Proof. For fixed j, p, ω, σ, this is the fixed-policy scalar frontier of Ariu et al. (2025) with target policy $\pi^{(j)}$, reference kernel p, product-kernel variable q, positive row weights ω, and weighted-KL budget σ. Their certified deterministic frontier routine returns $u_j^{\xi}(\sigma)$ with $u_j^{\xi}(\sigma) \geq u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p) \geq u_j^{\xi}(\sigma) - \xi^{\mathrm{stop}}$. Set $U_j(\sigma) := u_j^{\xi}(\sigma)$ and $L_j(\sigma) := u_j^{\xi}(\sigma) - \xi^{\mathrm{stop}}$. This imports only the deterministic optimization certificate; the online validity of $(\hat{p}_t, \hat{w}(t))$ is proved separately in Appendix J. □

Proposition 3.5 (Policywise certified stopping implication)**.** *Under the certified bracket of Proposition 3.4, if $L_j(\sigma) > 0$, then $u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p) > 0$. Consequently, every kernel $q \in \mathcal{P}$ satisfying $\Phi(p;\omega,q) \leq \sigma$ has the same j-th sign as p.*

Proof. The first claim follows from $L_j(\sigma) \leq u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p)$. If $u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p) > 0$, then by definition every feasible q obeys $V_{p,j}(\rho)V_{q,j}(\rho) > 0$. □

At round t, PT-ACE(μ) calls this oracle for every $j \in [J]$ with $(p,\omega,\sigma) = (\hat{p}_t, \hat{w}(t), \sigma_t)$, where $\sigma_t := \beta(t,\delta)/(t\vee 1)$. It receives brackets

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

The joint stopping rule is $\min_{j\in[J]} L_{t,j}^{\mathrm{stop}} > 0$, in which case the algorithm returns $(\mathrm{Ans}_1(\hat{p}_t), \ldots, \mathrm{Ans}_J(\hat{p}_t))$. On the standard self-normalized KL event, the true kernel p belongs to the plausible set at every time; hence if all lower certificates are positive, every plausible kernel has the same sign vector as $\hat{p}_t$, and in particular so does p. The formal PAC proof is given in Appendix J.

3.6. Main guarantee

We now combine the three components above: exchange learns a bottleneck allocation, navigation realizes it along the single trajectory, and the certified stopping interface turns the accumulated evidence into a joint sign certificate.

Assumption 3.6 (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.

Assumption 3.6 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.

The floor parameter μ is an upper-bound regularization. For the true instance p, the theorem is proved for sufficiently small admissible floors, so that the fixed-floor occupancy domains used by the proof are nonempty on the tail event. The algorithm itself can be run with any floor satisfying $|\mathcal{S}||\mathcal{A}|\mu \le 1$, but the guarantee below is for floors in the instance-dependent admissible range. The lower-bound benchmark $T^\star(p)$ has no fixed-floor restriction.

At an active exchange round, the cut routine is queried with $(\bar{p}_t, x_t, \xi_t^{\mathrm{cut}})$, where $x_t \in \widetilde{\Sigma}^{\mu}_{\bar{p}_t}$. It returns a bottleneck index j_t, a sign-flipped witness q_t, and the rowwise KL vector $g_t(s,a) := \mathrm{KL}(\bar{p}_t(\cdot \mid s,a) \| q_t(\cdot \mid s,a))$, satisfying

$$F^{1:J}_{\bar{p}_t}(x) - F^{1:J}_{\bar{p}_t}(x_t) \le \langle g_t, x - x_t\rangle + \xi_t^{\mathrm{cut}} \qquad \forall x \in \widetilde{\Sigma}^{\mu}_{\bar{p}_t}.$$

The stopping routine is queried with tolerance ξ_t^{stop}, the certified width of the policywise bracket. The theorem is stated conditionally on deterministic tolerances satisfying

$$\sup_{t\ge 0} \xi_t^{\mathrm{cut}} < \infty, \qquad \xi_t^{\mathrm{cut}} \to 0, \qquad \xi_t^{\mathrm{stop}} \to 0. \tag{4}$$

At every query, the cut and stopping routines return outputs satisfying the contracts above. These are numerical-interface requirements rather than statistical assumptions on the MDP; Appendices D and E explain how the policywise certified frontier solver of Ariu et al. (2025), together with witness extraction, instantiates them.

Theorem 3.7 (Asymptotic upper bound for PT-ACE(μ))**.** *Fix a sufficiently small admissible floor parameter $\mu > 0$. Suppose Assumptions 2.2 and 3.6 hold. Run PT-ACE(μ) with the certified stopping and cut subroutines described in Appendices D and E, using deterministic input tolerance schedules $(\xi_t^{\mathrm{cut}})_{t\ge0}$ and $(\xi_t^{\mathrm{stop}})_{t\ge0}$ satisfying (4). Then PT-ACE(μ) is δ-PC for joint testing and satisfies*

$$\limsup_{\delta\to 0} \frac{\mathbb{E}_p[\tau_\delta]}{\log(1/\delta)} \le (1 + c(\mu,p))T^\star(p),$$

where $c(\mu,p) \ge 0$ and $c(\mu,p) \to 0$ as $\mu \downarrow 0$. Consequently, $\{\mathrm{PT\text{-}ACE}(\mu)\}_{\mu>0}$ recovers the lower-bound characteristic time in the two-stage limit: first $\delta \downarrow 0$ for fixed admissible μ, and then $\mu \downarrow 0$.

The proof is given in Appendix J. It combines the three layers above: exchange learns a bottleneck allocation with frontier value at least $T^\star(p)^{-1}$ up to the factor $1 + c(\mu,p)$, navigation transfers this guarantee to the empirical allocation generated by the single trajectory, and the certified stopping interface turns the resulting frontier margin into a joint sign certificate.

4. Conclusion

We studied fixed-confidence joint policy testing in discounted tabular MDPs under active exploration from a single adaptive trajectory. The characteristic time $T^\star(p)$ gives the instance-specific first-order benchmark. We developed PT-ACE(μ), which combines stabilized exchange-based allocation learning, trajectory-compatible navigation, and certified policywise stopping. For every sufficiently small admissible fixed floor $\mu > 0$, PT-ACE(μ) matches this benchmark up to the factor $1 + c(\mu,p)$, with $c(\mu,p) \to 0$ as $\mu \downarrow 0$.

The main contribution is the exchange–navigation layer that makes the lower-bound allocation achievable online: it learns a bottleneck occupancy through stabilized moving-domain updates, realizes it through a policy induced by averaged occupancies, and stops only when certified policywise scalar frontier tests verify all target-policy signs.

Limitations. The fixed floor μ is a proof-side regularization, not part of the lower bound. Our upper-bound proof requires μ to be small enough for the true instance p, so that transported interior benchmarks remain feasible for the moving fixed-floor domains on the tail event. Hence the admissible range of μ is instance-dependent and unknown a priori. For any floor in this range, PT-ACE(μ) satisfies the stated asymptotic guarantee; taking such floors to zero recovers the lower-bound rate.

Future directions. A natural next step is a fully adaptive floor schedule that removes the need to know an admissible range. Other directions include removing the μ-conservative slack, proving optimality for an adaptive vanishing-floor algorithm, weakening the navigation-side ergodicity assumption, sharpening finite-δ guarantees, improving certified frontier implementations, and extending

the exchange–navigation–certification template to broader pure-exploration problems in MDPs.

Broader impact. This work is methodological: it provides tools for certifying the signs of several fixed policy values from one adaptive trajectory. Such certificates may support evidence-based decisions in applications such as recommendation, operations, and treatment-strategy design. However, our guarantees concern only the reward, target policies, and model class specified by the user; they do not by themselves ensure social desirability, fairness, or robustness to unmodeled shifts. High-stakes uses should therefore be paired with domain validation, safety constraints, and human oversight.

Acknowledgements

Po-An Wang is supported by NSTC Grant Number 114-2118-M-007 -002 -MY2. Kaito Ariu was supported by JST BOOST Program Grant Number JPMJBY25A0 and JSPS KAKENHI Grant Number JP25K21291.

References

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

Ariu, K., Wang, P.-A., Proutière, A., and Abe, K. (2025). Policy testing in markov decision processes. *arXiv preprint arXiv:2505.15342*.

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.

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

Kaufmann, E., Cappé, O., and Garivier, A. (2016). On the complexity of best-arm identification in multi-armed bandit models. *The 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 *International Conference on Machine Learning*, pages 1690–1698. PMLR.

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

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

Wang, P.-A., Tzeng, R.-C., and Proutiere, A. (2021). Fast pure exploration via frank-wolfe. *Advances in Neural Information Processing Systems*, 34:5810–5821.

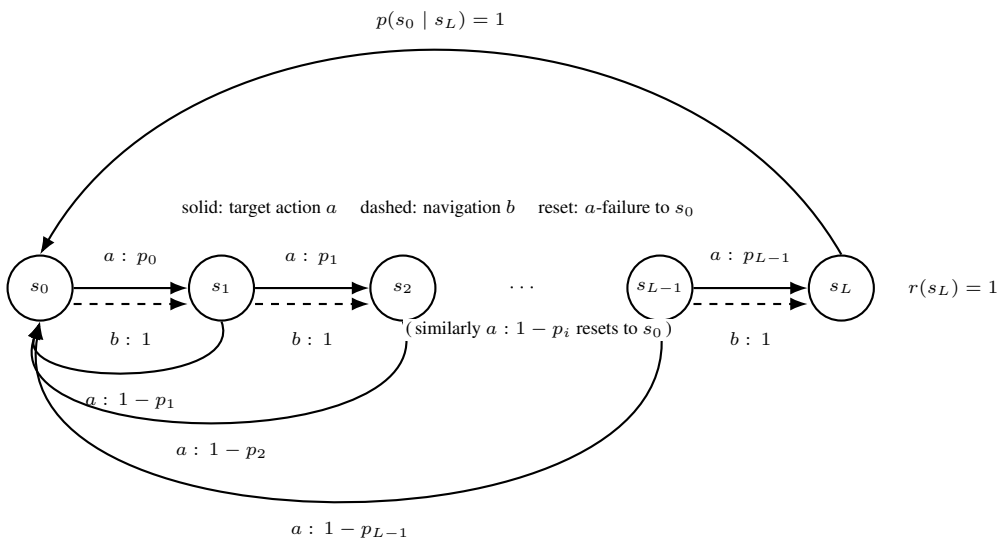

Figure 1. The reset-chain instance. State s_L is a rewarding reset state: it yields reward 1 and then deterministically returns to s_0.

Contents

Appendix roadmap

The appendices are organized as follows. Appendix A gives a reset-chain example showing why on-policy testing can be exponentially inefficient. Appendix B proves the sample-complexity lower bound. Appendix C records the policywise scalar-frontier facts used by the stopping interface. Appendix D explains how the certified policywise stopping oracle is instantiated from the fixed-policy testing solver of Ariu et al. (2025). Appendix E describes certified policywise thresholding and witness extraction for the exchange cut oracle. Appendix F proves adaptive kernel concentration. Appendix G analyzes the moving-domain exchange layer. Appendix H proves the navigation transfer from averaged occupancies to empirical frequencies. Appendix I proves the stopping-frontier sensitivity result on a compact interior slice. Appendix J proves the main upper bound. Finally, Appendix K collects frequently used notation.

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

This appendix gives a simple discounted tabular MDP illustrating why policy testing from an on-policy trajectory can be exponentially slower than active exploration. The example is not used in the proof of the main theorem. Its role is only to show that the learner may need to navigate to informative state-action pairs that the target policy rarely visits.

MDP instance. Fix $\gamma \in (0, 1)$. Let

$$\mathcal{S} = \{s_0, s_1, \ldots, s_L\}, \qquad \rho = \delta_{s_0}.$$

State s_L is a rewarding reset state: it yields reward 1 and then returns deterministically to s_0. All other rewards are zero: $r(s_L, \cdot) = 1$ and $r(s_i, \cdot) = 0$ for $i < L$.

For each $i = 0, \ldots, L-1$, there are two actions at s_i: a target action a and a navigation action b. The transitions are

$$p(s_{i+1} \mid s_i, a) = p_i, \qquad p(s_0 \mid s_i, a) = 1 - p_i, \qquad p(s_{i+1} \mid s_i, b) = 1.$$

At s_L, there is a single action and $p(s_0 \mid s_L) = 1$. The target policy π is deterministic: $\pi(s_i) = a$ for $i = 0, \ldots, L-1$, and $\pi(s_L)$ is the only available action.

Value under π. Under the target policy, a reward is obtained exactly when the trajectory completes a full successful run from s_0 to s_L.

Lemma A.1 (Value of the target policy on the reset chain)**.** *Let $\alpha_i := \prod_{j=0}^{i-1} p_j$, $i = 0, \dots, L$, with the convention $\alpha_0 = 1$. Then*

$$V_p^\pi(s_0) = \frac{\gamma^L \prod_{i=0}^{L-1} p_i}{1 - \left(\sum_{i=0}^{L-1} \gamma^{i+1}\alpha_i(1-p_i) + \gamma^{L+1}\prod_{i=0}^{L-1} p_i\right)}.$$

In particular,

$$\gamma^L \prod_{i=0}^{L-1} p_i \le V_p^\pi(s_0) \le \frac{\gamma^L}{1-\gamma}\prod_{i=0}^{L-1} p_i.$$

Proof. Let R be the discounted reward accumulated in one regenerative cycle from s_0 until the next return to s_0. Under π, the cycle either fails at some level $i < L$, returning to s_0 after $i+1$ steps, or succeeds all the way to s_L, receives reward 1 at time L, and returns to s_0 one step later. Hence

$$\mathbb{E}[R] = \gamma^L \prod_{i=0}^{L-1} p_i.$$

The discounted renewal factor is

$$\sum_{i=0}^{L-1} \gamma^{i+1}\alpha_i(1-p_i) + \gamma^{L+1}\prod_{i=0}^{L-1} p_i.$$

The regenerative identity $V_p^\pi(s_0) = \mathbb{E}[R] + \mathbb{E}[\gamma^{\tau_{\text{reset}}}]V_p^\pi(s_0)$ gives the formula. The lower bound follows because the denominator is at most 1, and the upper bound follows because the denominator is at least $1-\gamma$. □

Testing formulation. Fix a threshold $r_\star > 0$, and define the shifted reward $\tilde{r}(s,a) := r(s,a) - (1-\gamma)r_\star$. Then $V_{p,\tilde{r}}^\pi(s_0) = V_p^\pi(s_0) - r_\star$, so deciding whether $V_p^\pi(s_0)$ is above or below $r_\star$ is equivalent to deciding the sign of $V_{p,\tilde{r}}^\pi(s_0)$.

Lemma A.2 (On-policy occupation of the bottleneck row)**.** *Consider the reset chain under the target policy π, and suppose $p_i = 1/2$ for $i = 0, \dots, L-1$. Let*

$$N_{L-1}(T) := \sum_{t=0}^{T-1} \mathbf{1}\{S_t = s_{L-1}, A_t = a\}$$

be the number of on-policy visits to the bottleneck row before a stopping time T. Then

$$\mathbb{E}[N_{L-1}(T)] \le 2^{-(L-1)}\big(\mathbb{E}[T] + 1\big).$$

Proof. Decompose the on-policy trajectory into regenerative cycles, each starting at s_0 and ending at the next return to s_0. Let R_k be the indicator that the k-th cycle visits s_{L-1}. Since visiting s_{L-1} requires $L-1$ consecutive successful transitions from s_0, the variables R_k are i.i.d. Bernoulli with mean $2^{-(L-1)}$.

Let K_T be the number of cycles started before time T. The event $\{K_T \ge k\}$ is determined by the cycles before the k-th cycle starts, and is therefore independent of R_k. Also, $K_T \le T+1$, because every cycle lasts at least one transition. Since each cycle contains at most one visit to s_{L-1},

$$N_{L-1}(T) \le \sum_{k=1}^{K_T} R_k.$$

Using the predictability of $\{K_T \ge k\}$,

$$\mathbb{E}\left[\sum_{k=1}^{K_T} R_k\right] = \sum_{k\ge1} \mathbb{E}[\mathbf{1}\{K_T \ge k\}R_k] = 2^{-(L-1)}\sum_{k\ge1}\mathbb{P}(K_T \ge k) = 2^{-(L-1)}\mathbb{E}[K_T].$$

Since $K_T \le T+1$, the claim follows. □

Proposition A.3 (Two-point exponential–polynomial separation)**.** *Fix $\gamma \in (0,1)$ and $L \ge 1$. There exist two reset-chain instances p^+ and p^-, and a threshold $r_\star$, such that*

$$V_{p^+}^\pi(s_0) > r_\star > V_{p^-}^\pi(s_0).$$

Moreover, there exist constants $c_\gamma, C_\gamma > 0$, depending only on γ, such that the following hold.

(i) Any δ-PC testing strategy that is restricted to executing π, and that is correct on both p^+ and p^-, must satisfy

$$\mathbb{E}_{p^+}[T] \geq c_\gamma\, 2^L \log(1/\delta) - 1.$$

In particular, for all sufficiently small δ,

$$\mathbb{E}_{p^+}[T] \geq \frac{c_\gamma}{2}\, 2^L \log(1/\delta).$$

(ii) There is an active two-point testing strategy, correct on both p^+ and p^- with error probability at most δ, whose stopping time satisfies

$$\mathbb{E}_{p^+}[T] \vee \mathbb{E}_{p^-}[T] \leq C_\gamma\, L^2 \log(1/\delta).$$

Proof. Set

$$p_i^+ = \frac{1}{2}, \qquad i = 0, \ldots, L-1,$$

and let p^- differ from p^+ only at the last bottleneck:

$$p_i^- = p_i^+ \quad (i < L-1), \qquad p_{L-1}^- = \frac{1-\gamma}{8}.$$

Since $p_{L-1}^+ = 1/2$, we have

$$\prod_{i=0}^{L-1} p_i^- = \frac{1-\gamma}{4} \prod_{i=0}^{L-1} p_i^+.$$

By Lemma A.1,

$$V_{p^-}^\pi(s_0) \leq \frac{1}{4} V_{p^+}^\pi(s_0).$$

Choose

$$r_\star := \frac{1}{2}\big(V_{p^+}^\pi(s_0) + V_{p^-}^\pi(s_0)\big).$$

Then p^+ and p^- have opposite correct testing answers.

We first prove the on-policy lower bound. Under on-policy execution, the only statistical difference between p^+ and p^- is the transition at (s_{L-1}, a). Each visit to this row contributes

$$D_\gamma := \mathrm{KL}(\mathrm{Bern}(1/2) \,\|\, \mathrm{Bern}((1-\gamma)/8)) < \infty$$

to the expected log-likelihood ratio. Therefore, for any on-policy stopping rule,

$$\mathbb{E}_{p^+}\left[\log \frac{d\mathbb{P}_{p^+}^T}{d\mathbb{P}_{p^-}^T}\right] = D_\gamma\, \mathbb{E}_{p^+}[N_{L-1}(T)].$$

By Lemma A.2,

$$\mathbb{E}_{p^+}[N_{L-1}(T)] \leq 2^{-(L-1)}\big(\mathbb{E}_{p^+}[T] + 1\big).$$

A standard binary change-of-measure inequality for a δ-PC test between p^+ and p^- gives

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

Since $\mathrm{kl}(1-\delta, \delta) \geq c \log(1/\delta)$ for $\delta \in (0, 1/2)$, the lower bound follows after adjusting the constant c_γ.

We now give an active strategy. The two instances differ only in the Bernoulli transition at (s_{L-1}, a), where the success probabilities are $1/2$ under p^+ and $(1-\gamma)/8$ under p^-. To collect one probe of this transition, the learner navigates deterministically from s_0 to s_{L-1} using action b, then plays action a once at s_{L-1} and records whether the next state is s_L. After the probe, if the process is not already at s_0, the learner navigates to s_L using action b and takes the reset to s_0. Thus each probe costs at most $L+1$ transitions.

Let $n = \lceil C'_\gamma \log(1/\delta) \rceil$, where C'_γ is a sufficiently large constant depending only on γ. Perform n probes and let $\widehat{p}$ be the empirical success frequency. Since the two success probabilities are separated by the constant $1/2 - (1-\gamma)/8$, standard Chernoff bounds imply that the threshold test

$$\widehat{p} \gtrless \frac{1}{2}\left(\frac{1}{2} + \frac{1-\gamma}{8}\right)$$

distinguishes p^+ from p^- with error probability at most δ. The total number of transitions is at most

$$(L+1)\left\lceil C'_\gamma \log(1/\delta) \right\rceil \le C_\gamma L^2 \log(1/\delta),$$

after enlarging C_γ. This proves the active upper bound. □

Remark A.4 (Why this appendix is only illustrative). The active upper bound above is deliberately not optimized. Its purpose is only to show that active navigation can turn an exponentially rare on-policy informative event into direct probes of the relevant transition row. The main results of the paper provide the instance-specific asymptotic analysis for the general simultaneous testing problem.

B. Sample Complexity Lower Bound for Simultaneous Testing

This appendix proves Theorem 2.3. The proof is the simultaneous-testing analogue of the standard change-of-measure lower bound for online pure exploration in MDPs. The only point requiring some care is the alternative set. The δ-PC guarantee is imposed only on $\mathcal{P}_{\text{test}}^{1:J}$, so the direct change-of-measure argument can use only alternatives in this testing instance class. We first prove the lower bound with these testing-class alternatives and then show that this restriction does not change the characteristic-time objective used in the main text.

Introduce the visit counts up to round t:

$$N_s(t) := \sum_{i=1}^{t} \mathbf{1}\{s(i) = s\}, \qquad N_{sa}(t) := \sum_{i=1}^{t} \mathbf{1}\{s(i) = s,\ a(i) = a\}.$$

Recall that

$$\Pi_{\text{test}} := \{\pi^{(1)}, \dots, \pi^{(J)}\}, \qquad V_{p,j}(\rho) := V_p^{\pi^{(j)}}(\rho),$$

and

$$\mathcal{P}_{\text{test}}^{1:J} := \{p \in \mathcal{P} :\ V_{p,j}(\rho) \neq 0\ \forall j \in [J]\}.$$

For $p \in \mathcal{P}_{\text{test}}^{1:J}$, the answer vector is

$$\text{Ans}_{1:J}(p) := (\text{Ans}_1(p), \dots, \text{Ans}_J(p)), \qquad \text{Ans}_j(p) := \text{sign}(V_{p,j}(\rho)).$$

For the change-of-measure proof, define the policywise testing-class alternatives

$$\text{Alt}_j^{\text{test}}(p) := \{q \in \mathcal{P}_{\text{test}}^{1:J} :\ V_{q,j}(\rho) V_{p,j}(\rho) < 0\},$$

and

$$\text{Alt}_{1:J}^{\text{test}}(p) := \bigcup_{j=1}^{J} \text{Alt}_j^{\text{test}}(p).$$

Equivalently, $q \in \text{Alt}_{1:J}^{\text{test}}(p)$ if and only if $q \in \mathcal{P}_{\text{test}}^{1:J}$ and $\text{Ans}_{1:J}(q) \neq \text{Ans}_{1:J}(p)$.

We also keep the main-text alternative set

$$\text{Alt}_j(p) := \{q \in \mathcal{P} :\ V_{q,j}(\rho) V_{p,j}(\rho) < 0\}, \qquad \text{Alt}_{1:J}(p) := \bigcup_{j=1}^{J} \text{Alt}_j(p).$$

The set $\text{Alt}_{1:J}(p)$ may contain kernels outside $\mathcal{P}_{\text{test}}^{1:J}$, for instance if some other target policy has zero value under q. Lemma B.3 below shows that this distinction does not affect the characteristic-time objective.

As in the main text, for an allocation ω and a kernel q, write

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

Lemma B.1 (Al Marjani et al., 2021)**.** *Let $p \in \mathcal{P}$. For any strategy and any state $y \in \mathcal{S}$, one has*

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

Lemma B.2 (Information constraint for simultaneous policy testing)**.** *For any δ-PC simultaneous-testing algorithm, any $p \in \mathcal{P}_{\text{test}}^{1:J}$, and any $q \in \text{Alt}_{1:J}^{\text{test}}(p)$, one has*

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

Proof. Fix $p \in \mathcal{P}_{\text{test}}^{1:J}$ and $q \in \text{Alt}_{1:J}^{\text{test}}(p)$. Define the $\mathcal{F}_\tau$-measurable event

$$E_p := \{\hat{\imath} = \text{Ans}_{1:J}(p)\}.$$

Since the algorithm is δ-PC for simultaneous testing,

$$\mathbb{P}_p(E_p) \geq 1 - \delta.$$

Because $q \in \text{Alt}_{1:J}^{\text{test}}(p)$, we have $\text{Ans}_{1:J}(q) \neq \text{Ans}_{1:J}(p)$. Hence, on E_p, the returned sign vector is wrong under q. Since $q \in \mathcal{P}_{\text{test}}^{1:J}$, the δ-PC guarantee also applies under q, and therefore

$$\mathbb{P}_q(E_p) \leq \delta.$$

Applying the standard change-of-measure inequality for adaptive trajectories to E_p yields

$$\sum_{s,a} \mathbb{E}_p[N_{sa}(\tau)]\, \text{KL}_{sa}(p, q) \geq \text{kl}\big(\mathbb{P}_p(E_p), \mathbb{P}_q(E_p)\big).$$

Using $\mathbb{P}_p(E_p) \geq 1 - \delta$, $\mathbb{P}_q(E_p) \leq \delta$, and the monotonicity of binary relative entropy gives

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

□

Lemma B.3 (Testing-class and support-safe reduction)**.** *Fix $p \in \mathcal{P}_{\text{test}}^{1:J}$. For every $\omega \in \Delta(\mathcal{S} \times \mathcal{A})$,*

$$\inf_{q \in \text{Alt}_{1:J}(p)} \Phi(p; \omega, q) = \inf_{q \in \text{Alt}_{1:J}^{\text{test}}(p)} \Phi(p; \omega, q) = \inf_{q \in \text{Alt}_{1:J}^{\text{test},\circ}(p)} \Phi(p; \omega, q),$$

where

$$\text{Alt}_{1:J}^{\text{test},\circ}(p) := \{q \in \text{Alt}_{1:J}^{\text{test}}(p) : \text{KL}_{sa}(p, q) < \infty \ \forall (s, a)\}.$$

Proof. Since

$$\text{Alt}_{1:J}^{\text{test},\circ}(p) \subseteq \text{Alt}_{1:J}^{\text{test}}(p) \subseteq \text{Alt}_{1:J}(p),$$

one inequality is immediate. It remains to approximate every finite-value alternative in $\text{Alt}_{1:J}(p)$ by alternatives in $\text{Alt}_{1:J}^{\text{test},\circ}(p)$.

Fix $q \in \text{Alt}_{1:J}(p)$ with $\Phi(p; \omega, q) < \infty$. Then for some $j_0 \in [J]$,

$$V_{q,j_0}(\rho) V_{p,j_0}(\rho) < 0.$$

By Assumption 2.2, the strict j_0-opposite-sign region is nonempty. We claim that it contains a full-row-support kernel q° such that

$$V_{q^\circ,k}(\rho) \neq 0 \qquad \forall k \in [J].$$

To prove the claim, let

$$\mathcal{P}^\circ := \{q' \in \mathcal{P} : q'(y \mid s, a) > 0 \ \forall (s, a, y)\}$$

be the relative interior of the product simplex. It is connected and relatively open. For each $k \in [J]$, the map

$$q' \mapsto V_{q',k}(\rho)$$

is real analytic on $\mathcal{P}^\circ$, since

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

and $P_{q'}^{\pi^{(k)}}$ depends affinely on q'. This analytic function is not identically zero on $\mathcal{P}^\circ$: by Assumption 2.2, the k-th value can be made strictly positive and strictly negative by suitable kernels, and mixing those kernels with an arbitrarily small amount of the uniform kernel moves them into $\mathcal{P}^\circ$ without changing the strict sign. Hence the zero set

$$\{q' \in \mathcal{P}^\circ : V_{q',k}(\rho) = 0\}$$

has empty interior in $\mathcal{P}^\circ$. Since $J < \infty$, the union of these zero sets also has empty interior.

The strict j_0-opposite-sign region is open. Since it is nonempty and full-row-support kernels are dense, its intersection with $\mathcal{P}^\circ$ is a nonempty open subset of $\mathcal{P}^\circ$. Therefore it contains a point outside the finite union of the zero sets above. This point is the desired q°.

For $\varepsilon \in (0,1)$, set

$$q_\varepsilon := (1-\varepsilon)q + \varepsilon q^\circ.$$

Then q_ε has full row support. By continuity, for all sufficiently small $\varepsilon > 0$,

$$V_{q_\varepsilon,j_0}(\rho)V_{p,j_0}(\rho) < 0.$$

Moreover, for each k, the map $\varepsilon \mapsto V_{q_\varepsilon,k}(\rho)$ is real analytic and is not identically zero, since its value at $\varepsilon = 1$ is $V_{q^\circ,k}(\rho) \neq 0$. Hence its zeros are isolated. Since $J < \infty$, there exists a sequence $\varepsilon_m \downarrow 0$ such that

$$V_{q_{\varepsilon_m},k}(\rho) \neq 0 \qquad \forall k \in [J],\ \forall m.$$

Thus

$$q_{\varepsilon_m} \in \mathrm{Alt}^{\mathrm{test},\circ}_{1:J}(p).$$

Finally, because $\Phi(p;\omega,q) < \infty$, every row with $\omega_{sa} > 0$ is support-safe for the KL term. Since $q_{\varepsilon_m} \to q$ rowwise and $q_{\varepsilon_m} \geq (1-\varepsilon_m)q$, the rowwise KL terms converge on all rows with $\omega_{sa} > 0$; rows with $\omega_{sa} = 0$ do not contribute. Therefore

$$\Phi(p;\omega,q_{\varepsilon_m}) \to \Phi(p;\omega,q).$$

Taking infima gives the reverse inequality and proves the claim. □

For $\delta \in (0,1/2)$, define $k_\delta := \mathrm{kl}(\delta, 1-\delta)$. Let $v_\delta^{1:J}(p)$ denote the value of the finite-δ count program

$$\begin{aligned} v_\delta^{1:J}(p) := \inf_{n\geq 0} \quad & \sum_{s,a} n_{sa} \\ \text{s.t.} \quad & \forall y \in \mathcal{S}, \quad \left| \sum_a n_{ya} - \sum_{s,a} p(y \mid s,a)\, n_{sa} \right| \leq 1, \\ & \forall q \in \mathrm{Alt}^{\mathrm{test},\circ}_{1:J}(p), \quad \sum_{s,a} n_{sa}\mathrm{KL}_{sa}(p,q) \geq k_\delta. \end{aligned} \tag{5}$$

Proposition B.4. *For any δ-PC simultaneous-testing strategy,*

$$\liminf_{\delta\to 0} \frac{\mathbb{E}_p[\tau]}{\log(1/\delta)}$$

is lower bounded by the value of the following optimization problem:

$$\begin{aligned} \inf_{n\geq 0} \quad & \sum_{s,a} n_{sa} \\ \textit{s.t.} \quad & \forall y \in \mathcal{S}, \quad \sum_a n_{ya} = \sum_{s,a} p(y \mid s,a)\, n_{sa}, \\ & \forall q \in \mathrm{Alt}^{\mathrm{test},\circ}_{1:J}(p), \quad \sum_{s,a} n_{sa}\mathrm{KL}_{sa}(p,q) \geq 1. \end{aligned} \tag{6}$$

Proof. By Lemma B.1 and Lemma B.2, every δ-PC simultaneous-testing strategy satisfies

$$\mathbb{E}_p[\tau] \geq v_\delta^{1:J}(p),$$

where $v_\delta^{1:J}(p)$ is the value of (5). Hence it is enough to show that

$$\liminf_{\delta\to 0} \frac{v_\delta^{1:J}(p)}{\log(1/\delta)}$$

is lower bounded by the value of (6).

Let $\{\delta_m\}_{m\geq 1} \subset (0, 1/2)$ be any sequence such that $\delta_m \downarrow 0$, and for each m choose $n^{(m)}$ feasible for (5) at level δ_m such that

$$\sum_{s,a} n_{sa}^{(m)} \leq v_{\delta_m}^{1:J}(p) + 1.$$

Define $\bar{n}^{(m)} := n^{(m)}/k_{\delta_m}$. Since $k_{\delta_m} \to \infty$, the feasibility conditions for $n^{(m)}$ imply

$$\left| \sum_a \bar{n}_{ya}^{(m)} - \sum_{s,a} p(y \mid s, a)\, \bar{n}_{sa}^{(m)} \right| \leq \frac{1}{k_{\delta_m}} \quad \forall y \in \mathcal{S},$$

and

$$\sum_{s,a} \bar{n}_{sa}^{(m)} \mathrm{KL}_{sa}(p, q) \geq 1 \quad \forall q \in \mathrm{Alt}_{1:J}^{\mathrm{test},\circ}(p).$$

Now fix a subsequence, still denoted by m, along which

$$\sum_{s,a} \bar{n}_{sa}^{(m)} \to \liminf_{m\to\infty} \sum_{s,a} \bar{n}_{sa}^{(m)}.$$

If this limit is $+\infty$, there is nothing to prove. Otherwise, $(\bar{n}^{(m)})_{m\geq 1}$ is bounded in the finite-dimensional space $\mathbb{R}_+^{\mathcal{S}\times\mathcal{A}}$. After extraction, assume $\bar{n}^{(m)} \to \bar{n} \in \mathbb{R}_+^{\mathcal{S}\times\mathcal{A}}$.

Passing to the limit in the flow constraints gives

$$\sum_a \bar{n}_{ya} = \sum_{s,a} p(y \mid s, a)\, \bar{n}_{sa} \qquad \forall y \in \mathcal{S}.$$

For each fixed $q \in \mathrm{Alt}_{1:J}^{\mathrm{test},\circ}(p)$, all rowwise divergences $\mathrm{KL}_{sa}(p, q)$ are finite constants. Hence

$$\sum_{s,a} \bar{n}_{sa} \mathrm{KL}_{sa}(p, q) = \lim_{m\to\infty} \sum_{s,a} \bar{n}_{sa}^{(m)} \mathrm{KL}_{sa}(p, q) \geq 1.$$

Thus $\bar{n}$ is feasible for (6). Therefore

$$\liminf_{m\to\infty} \sum_{s,a} \bar{n}_{sa}^{(m)} \geq \mathrm{val}(6),$$

where $\mathrm{val}(6)$ denotes the value of (6). Since

$$\sum_{s,a} \bar{n}_{sa}^{(m)} = \frac{\sum_{s,a} n_{sa}^{(m)}}{k_{\delta_m}} \leq \frac{v_{\delta_m}^{1:J}(p) + 1}{k_{\delta_m}},$$

we obtain

$$\liminf_{\delta\to 0} \frac{v_\delta^{1:J}(p)}{k_\delta} \geq \mathrm{val}(6).$$

Finally, $k_\delta / \log(1/\delta) \to 1$ as $\delta \to 0$, so

$$\liminf_{\delta\to 0} \frac{v_\delta^{1:J}(p)}{\log(1/\delta)} \geq \mathrm{val}(6).$$

This proves the proposition. □

Proof of Theorem 2.3. By Proposition B.4, it suffices to show that the value of (6) is at least

$$\left(\max_{\omega\in\Sigma_p}\inf_{q\in\mathrm{Alt}_{1:J}^{\mathrm{test},\circ}(p)}\sum_{s,a}\omega_{sa}\mathrm{KL}_{sa}(p,q)\right)^{-1}.$$

By Lemma B.3, this is equal to

$$\left(\max_{\omega\in\Sigma_p}\inf_{q\in\mathrm{Alt}_{1:J}(p)}\sum_{s,a}\omega_{sa}\mathrm{KL}_{sa}(p,q)\right)^{-1}=T^\star(p).$$

Let $\mathcal{N}_{1:J}$ denote the feasible set of (6):

$$\mathcal{N}_{1:J}:=\left\{n\in\mathbb{R}_+^{\mathcal{S}\times\mathcal{A}}:\ \begin{array}{ll}\sum_a n_{ya}=\sum_{s,a}p(y\mid s,a)n_{sa} & \forall y\in\mathcal{S},\\ \sum_{s,a}n_{sa}\mathrm{KL}_{sa}(p,q)\geq 1 & \forall q\in\mathrm{Alt}_{1:J}^{\mathrm{test},\circ}(p)\end{array}\right\}.$$

Then $\mathrm{val}(6)=\inf_{n\in\mathcal{N}_{1:J}}\sum_{s,a}n_{sa}$.

Fix any $n\in\mathcal{N}_{1:J}$, and let $M:=\sum_{s,a}n_{sa}$. Since $n\neq 0$ by the KL constraints, $M>0$. Define $\omega_{sa}:=n_{sa}/M$. Then $\omega\geq 0$, $\sum_{s,a}\omega_{sa}=1$, and for every $y\in\mathcal{S}$,

$$\sum_a\omega_{ya}=\frac{1}{M}\sum_a n_{ya}=\frac{1}{M}\sum_{s,a}p(y\mid s,a)n_{sa}=\sum_{s,a}p(y\mid s,a)\omega_{sa}.$$

Hence $\omega\in\Sigma_p$. Moreover, for every $q\in\mathrm{Alt}_{1:J}^{\mathrm{test},\circ}(p)$,

$$\sum_{s,a}\omega_{sa}\mathrm{KL}_{sa}(p,q)=\frac{1}{M}\sum_{s,a}n_{sa}\mathrm{KL}_{sa}(p,q)\geq\frac{1}{M}.$$

Therefore

$$\inf_{q\in\mathrm{Alt}_{1:J}^{\mathrm{test},\circ}(p)}\sum_{s,a}\omega_{sa}\mathrm{KL}_{sa}(p,q)\geq\frac{1}{M}.$$

Since $\omega\in\Sigma_p$, it follows that

$$\max_{\omega'\in\Sigma_p}\inf_{q\in\mathrm{Alt}_{1:J}^{\mathrm{test},\circ}(p)}\sum_{s,a}\omega'_{sa}\mathrm{KL}_{sa}(p,q)\geq\frac{1}{M}.$$

Equivalently,

$$M\geq\left(\max_{\omega'\in\Sigma_p}\inf_{q\in\mathrm{Alt}_{1:J}^{\mathrm{test},\circ}(p)}\sum_{s,a}\omega'_{sa}\mathrm{KL}_{sa}(p,q)\right)^{-1}.$$

Since this holds for every $n\in\mathcal{N}_{1:J}$, taking the infimum over $n\in\mathcal{N}_{1:J}$ yields

$$\mathrm{val}(6)\geq\left(\max_{\omega\in\Sigma_p}\inf_{q\in\mathrm{Alt}_{1:J}^{\mathrm{test},\circ}(p)}\sum_{s,a}\omega_{sa}\mathrm{KL}_{sa}(p,q)\right)^{-1}.$$

Using Lemma B.3 once more gives

$$\mathrm{val}(6)\geq T^\star(p).$$

This proves Theorem 2.3.

C. Policywise scalar frontier facts

This appendix records the policywise scalar-frontier facts used by the stopping interface. These facts are the single-policy testing facts of Ariu et al. (2025), applied separately to each target policy $\pi^{(j)}$. We include the notation to make clear how the imported policywise oracle connects to the simultaneous-testing wrapper.

Throughout this appendix, fix a policy index $j \in [J]$, a model $p \in \mathcal{P}_{\text{test}}^{1:J}$, and an allocation $\omega \in \Delta(\mathcal{S} \times \mathcal{A})$ with $\omega_{sa} > 0$ for every (s, a). Define

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

The policywise no-opposite-sign frontier is

$$u_{\text{NO}}^{(j)}(\sigma; \omega, p) := \inf\{g_{p,j}(q) : h(q) \leq \sigma\}.$$

The policywise nonconvex information frontier is

$$\sigma_{\text{NC}}^{(j)}(u; \omega, p) := \inf\{h(q) : g_{p,j}(q) < u\},$$

and the sign-flipped frontier is

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

Proposition C.1 (Policywise scalar frontier duality)**.** *Assume $\omega_{sa} > 0$ for every (s, a), and let $p \in \mathcal{P}_{\text{test}}^{1:J}$. For each $j \in [J]$, the scalar frontiers $u_{\text{NO}}^{(j)}$ and $\sigma_{\text{NC}}^{(j)}$ are inverse on the nontrivial range:*

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

whenever

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

In particular,

$$\sigma_j^\star(\omega, p) = \inf\{\sigma \geq 0 : u_{\text{NO}}^{(j)}(\sigma; \omega, p) \leq 0\}.$$

Proof. This is exactly the policywise scalar frontier duality theorem for fixed-policy testing in Ariu et al. (2025), applied to the target policy $\pi^{(j)}$. The strict positivity of ω ensures that zero weighted KL radius identifies $q = p$ on all rows relevant to the frontier, and Assumption 2.2 ensures the zero level is in the nontrivial range for every policy in the finite family. □

Lemma C.2 (Policywise stopping implication)**.** *If*

$$u_{\text{NO}}^{(j)}(\sigma; \omega, p) > 0,$$

then every kernel $q \in \mathcal{P}$ satisfying

$$\Phi(p; \omega, q) \leq \sigma$$

has the same j-th sign as p. That is,

$$V_{p,j}(\rho)V_{q,j}(\rho) > 0.$$

Proof. By definition,

$$u_{\text{NO}}^{(j)}(\sigma; \omega, p) = \inf_{\Phi(p;\omega,q) \leq \sigma} V_{p,j}(\rho)V_{q,j}(\rho).$$

If this infimum is positive, then the displayed product is positive for every kernel in the plausible set. □

Role in the paper. The online algorithm does not need to localize $\sigma_j^\star(\omega, p)$ at every round. It only needs a certified sign decision for $u_{\text{NO}}^{(j)}$ at the current budget $(\omega, p, \sigma) = (\hat{w}(t), \hat{p}_t, \sigma_t)$. Proposition C.1 explains why this scalar sign test represents the same policywise testing frontier that appears in the lower bound, while Lemma C.2 is the direct implication used in the PAC proof.

D. Instantiating the imported certified policywise stopping oracle

This appendix explains how the policywise stopping queries used by PT-ACE(μ) are instantiated from the certified fixed-policy testing solver of Ariu et al. (2025). The use of this solver is at the deterministic optimization-interface level. We do not invoke the static-sampling statistical theorem of Ariu et al. (2025), nor do we reproduce the internal projected-gradient routine, its discretization, or its certification proof. Instead, we record the reduction showing that, for each policy index j, the query made by PT-ACE(μ) has the same fixed-policy frontier form as in that work. The validity of using the data-dependent pair $(\hat{p}_t, \hat{w}(t))$, produced by a single adaptive trajectory, is proved separately by the concentration, exchange, navigation, and PAC arguments of the present paper.

Fix a policy index $j \in [J]$, a reference kernel $p \in \mathcal{P}_{\text{test}}^{1:J}$, an allocation $\omega \in \Delta(\mathcal{S} \times \mathcal{A})$ with strictly positive coordinates, and a budget $\sigma \geq 0$. The policywise stopping frontier queried by PT-ACE(μ) is

$$u_{\text{NO}}^{(j)}(\sigma; \omega, p) := \inf\Big\{V_{p,j}(\rho)V_{q,j}(\rho) : q \in \mathcal{P},\ \Phi(p; \omega, q) \leq \sigma\Big\},$$

where

$$\Phi(p; \omega, q) := \sum_{s,a} \omega_{sa}\text{KL}_{sa}(p, q).$$

For fixed j, this is the fixed-policy testing frontier for the target policy $\pi^{(j)}$, with reference kernel p, sampling weights ω, and weighted-KL budget σ. The factor $V_{p,j}(\rho)$ is fixed during the query, so minimizing

$$q \longmapsto V_{p,j}(\rho)V_{q,j}(\rho)$$

is exactly the scalar no-opposite-sign frontier for the j-th target policy.

The optimization variable is the transition kernel q, equivalently a product of independent row variables

$$q \in \prod_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Delta(\mathcal{S}).$$

This is the same product-kernel decision variable used in the static policy-testing framework. The objective $q \mapsto V_{p,j}(\rho)V_{q,j}(\rho)$ is the policy-value product studied there: its sign records whether the candidate kernel q gives the same answer as the reference kernel p for the fixed target policy $\pi^{(j)}$.

The reversed-MDP construction of Ariu et al. (2025) applies directly to this fixed-j problem. In that construction, a row (s, a) of the original transition kernel becomes a state of the reversed MDP, the reversed action selects the next state $y \in \mathcal{S}$, and a reversed policy encodes the row $q(\cdot \mid s, a)$. Under this correspondence, optimizing over transition kernels q is equivalent to optimizing over policies in the reversed MDP, and the induced value equals $V_{q,j}(\rho)$, with the target policy $\pi^{(j)}$ and the reward for the j-th test fixed.

The constraint is also the same product-budget constraint:

$$\Phi(p; \omega, q) = \sum_{s,a} \omega_{sa}\text{KL}(p(\cdot \mid s, a) \,\|\, q(\cdot \mid s, a)) \leq \sigma.$$

It is a separable rowwise KL budget with positive row weights. In PT-ACE(μ), the stopping oracle is queried only after the initial coverage phase, so every empirical coordinate satisfies $\hat{w}_{sa}(t) > 0$. Thus, at each stopping query, the allocation passed to the policywise solver lies in the positive-weight regime required by the scalar-frontier and product-budget construction of Ariu et al. (2025).

There is one important distinction from the setting of Ariu et al. (2025). Their statistical theorem is stated for fixed-policy testing under a generative model with a static sampling allocation, whereas PT-ACE(μ) generates $(\hat{p}_t, \hat{w}(t))$ from a single adaptive trajectory. We therefore import only the deterministic certified optimization interface of their solver. Conditional on a reference kernel p, a strictly positive allocation ω, a budget σ, and a fixed target policy $\pi^{(j)}$, the optimization problem above is the fixed-policy frontier treated by Ariu et al. (2025). The single-trajectory validity of using $(p, \omega) = (\hat{p}_t, \hat{w}(t))$ is proved in our PAC and upper-bound arguments through the self-normalized KL event, the initial coverage phase, and the exchange–navigation analysis. If a target policy $\pi^{(j)}$ is not full support, the same deterministic reduction is applied on the support-reduced reversed MDP; rows and actions that never affect $V_{q,j}(\rho)$ may be fixed to the reference kernel.

We now state the precise bracket returned by the imported interface. Applying the certified fixed-policy testing solver of Ariu et al. (2025) to the target policy $\pi^{(j)}$, with reference kernel p, positive allocation ω, budget σ, and tolerance $\xi^{\mathrm{stop}} > 0$, returns a computable scalar $u_j^{\xi}(\sigma)$ satisfying

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

We therefore define

$$U_j(\sigma) := u_j^{\xi}(\sigma), \qquad L_j(\sigma) := u_j^{\xi}(\sigma) - \xi^{\mathrm{stop}}.$$

Then

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

This is the certified bracket contract stated in Proposition 3.4 and used through Proposition 3.5.

PT-ACE(μ) applies this deterministic interface independently to $j = 1, \ldots, J$. No joint nonconvex stopping problem is solved. The simultaneous wrapper simply requires all policywise lower certificates to be positive:

$$\min_{j \in [J]} L_j(\sigma) > 0.$$

When this holds, Proposition 3.5 implies that every kernel in the current weighted-KL plausible set has the same sign as the reference kernel for every target policy in the finite family.

Role in the proof. The upper-bound proof never uses the static-sampling optimality theorem of Ariu et al. (2025), nor does it use the internal details of their numerical solver. It uses only the certified bracket guarantee displayed above. The new online work in this paper is the exchange–navigation layer: learning a near-optimal bottleneck allocation, realizing it along a single trajectory, and showing that the resulting empirical allocation eventually gives all imported policywise frontier tests a positive margin.

E. Certified policywise thresholds and bottleneck cuts

This appendix records the deterministic numerical contract used to instantiate the bottleneck cut oracle required by the exchange layer. We first recall a certified bisection wrapper for localizing a policywise scalar threshold. We then state a separate certified witness contract: a policywise routine must return a near-minimizing sign-flipped kernel, not only a certified scalar bracket. Finally, we prove that this witness contract implies the inexact bottleneck-cut contract used by PT-ACE(μ).

This appendix should be read as a cut-oracle reduction. Appendix D instantiates the certified scalar stopping brackets from the fixed-policy testing solver of Ariu et al. (2025). The exchange layer needs the additional witness information stated in Definition E.2. The upper-bound theorem assumes PT-ACE(μ) is run with a certified bottleneck-cut subroutine satisfying this deterministic witness contract; the proof does not use the static-sampling statistical theorem of Ariu et al. (2025) for this step.

Throughout the appendix, fix a policy index $j \in [J]$, a model $p \in \mathcal{P}_{\text{test}}^{1:J}$, and an allocation $\omega \in \Delta(\mathcal{S} \times \mathcal{A})$ with strictly positive coordinates. Recall the policywise no-opposite-sign frontier

$$u_{\text{NO}}^{(j)}(\sigma; \omega, p) := \inf\Big\{V_{p,j}(\rho)V_{q,j}(\rho) : q \in \mathcal{P},\ \Phi(p; \omega, q) \le \sigma\Big\},$$

where

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

The associated policywise sign-flipped threshold is

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

Equivalently, by the scalar-frontier duality of Proposition C.1,

$$\sigma_j^\star(\omega, p) = \inf_{q \in \mathrm{Alt}_j(p)} \Phi(p; \omega, q),$$

and the same value is obtained if $\mathrm{Alt}_j(p)$ is replaced by the closed alternative set

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

E.1. Certified sign queries

By Proposition 3.5, for every query $(\sigma, \xi^{\text{stop}})$, the policywise certified solver returns computable scalars $L_j(\sigma)$ and $U_j(\sigma)$ such that

$$L_j(\sigma) \le u_{\text{NO}}^{(j)}(\sigma; \omega, p) \le U_j(\sigma), \qquad U_j(\sigma) - L_j(\sigma) \le \xi^{\text{stop}}. \tag{7}$$

Hence a strict sign is certified whenever the bracket lies entirely on one side of zero:

$$L_j(\sigma) > 0 \quad \Longrightarrow \quad u_{\text{NO}}^{(j)}(\sigma; \omega, p) > 0,$$

and

$$U_j(\sigma) < 0 \quad \Longrightarrow \quad u_{\text{NO}}^{(j)}(\sigma; \omega, p) < 0.$$

If $0 \in [L_j(\sigma), U_j(\sigma)]$, the sign is inconclusive at the current tolerance. In that case one may rerun the solver with a smaller tolerance. For definiteness, we use the update

$$\xi^{\text{stop}} \leftarrow \xi^{\text{stop}}/2.$$

We denote by

$$(\mathrm{sgn}, L, U, \xi^{\text{stop}}) \leftarrow \textsc{CertSign}_j(\sigma, \xi^{\text{stop}})$$

the routine that repeatedly applies this refinement until a strict sign is certified, returning $\mathrm{sgn} \in \{\mathsf{POS}, \mathsf{NEG}\}$.

Algorithm 2 Certified bisection for $\sigma_j^\star(\omega, p)$

Require: target accuracy $\epsilon > 0$, initial tolerance $\xi_{\text{stop}}^{(0)} > 0$, initial upper guess $\sigma_{\text{hi}}^{(0)} > 0$

$\sigma_{\text{lo}} \leftarrow 0$, $\sigma_{\text{hi}} \leftarrow \sigma_{\text{hi}}^{(0)}$, $\xi_{\text{stop}} \leftarrow \xi_{\text{stop}}^{(0)}$
Bracketing phase
while true **do**
  $(\text{sgn}, L, U, \xi_{\text{stop}}) \leftarrow \textsc{CertSign}_j(\sigma_{\text{hi}}, \xi_{\text{stop}})$
  if $\text{sgn} = \mathsf{NEG}$ **then**
    break
  end if
  $\sigma_{\text{hi}} \leftarrow 2\sigma_{\text{hi}}$
end while
Bisection phase
while $\sigma_{\text{hi}} - \sigma_{\text{lo}} > \epsilon$ **do**
  $\sigma \leftarrow (\sigma_{\text{lo}} + \sigma_{\text{hi}})/2$
  $(\text{sgn}, L, U, \xi_{\text{stop}}) \leftarrow \textsc{CertSign}_j(\sigma, \xi_{\text{stop}})$
  if $\text{sgn} = \mathsf{POS}$ **then**
    $\sigma_{\text{lo}} \leftarrow \sigma$
  else
    $\sigma_{\text{hi}} \leftarrow \sigma$
  end if
end while
return $\hat{\sigma} \leftarrow (\sigma_{\text{lo}} + \sigma_{\text{hi}})/2$

E.2. Certified bisection for the policywise threshold

Starting from an initial upper guess $\sigma_{\text{hi}}^{(0)} > 0$, we first certify a negative upper endpoint by repeated doubling. We then run standard bisection, except that the sign at each midpoint is provided by $\textsc{CertSign}_j$ rather than by exact access to $u_{\text{NO}}^{(j)}(\sigma; \omega, p)$.

Proposition E.1 (Certified recovery of $\sigma_j^\star(\omega, p)$)**.** *Let*

$$\sigma_j^\star(\omega, p) := \inf\{\sigma \geq 0 : \; u_{\text{NO}}^{(j)}(\sigma; \omega, p) \leq 0\}.$$

Assume that $\textsc{CertSign}_j$ *always returns a correct strict sign, obtained from a certified bracket satisfying* (7)*. Then Algorithm 2 maintains*

$$\sigma_j^\star(\omega, p) \in [\sigma_{\text{lo}}, \sigma_{\text{hi}}]$$

throughout the bisection phase and returns $\hat{\sigma}$ *such that*

$$|\hat{\sigma} - \sigma_j^\star(\omega, p)| \leq \epsilon.$$

Moreover, the number of bisection updates is at most

$$\left\lceil \log_2\left(\frac{\sigma_{\text{hi}}^{\text{init}} - \sigma_{\text{lo}}^{\text{init}}}{\epsilon}\right)\right\rceil,$$

where $[\sigma_{\text{lo}}^{\text{init}}, \sigma_{\text{hi}}^{\text{init}}]$ *denotes the certified bracket at the start of the bisection phase.*

Proof. Because the feasible set $\{q : \Phi(p; \omega, q) \leq \sigma\}$ expands with σ, the map $\sigma \mapsto u_{\text{NO}}^{(j)}(\sigma; \omega, p)$ is nonincreasing.

Whenever $\textsc{CertSign}_j$ returns $\text{sgn} = \mathsf{POS}$, we have $L_j(\sigma) > 0$, hence by (7),

$$u_{\text{NO}}^{(j)}(\sigma; \omega, p) > 0.$$

Likewise, whenever $\textsc{CertSign}_j$ returns $\text{sgn} = \mathsf{NEG}$, we have $U_j(\sigma) < 0$, hence

$$u_{\text{NO}}^{(j)}(\sigma; \omega, p) < 0.$$

Thus every strict sign used by the algorithm is correct.

The bracketing phase terminates only when a negative upper endpoint has been certified, so $u_{\mathrm{NO}}^{(j)}(\sigma_{\mathrm{hi}};\omega,p) < 0$. Since $u_{\mathrm{NO}}^{(j)}(\cdot;\omega,p)$ is nonincreasing, this implies

$$\sigma_j^\star(\omega,p) \le \sigma_{\mathrm{hi}}.$$

Also, since ω has strictly positive coordinates, $\Phi(p;\omega,q) = 0$ implies $q = p$. Therefore

$$u_{\mathrm{NO}}^{(j)}(0;\omega,p) = V_{p,j}(\rho)^2 > 0,$$

because $p \in \mathcal{P}_{\mathrm{test}}^{1:J}$. Hence $\sigma_{\mathrm{lo}} = 0$ is a valid lower endpoint.

Now let $\sigma = (\sigma_{\mathrm{lo}} + \sigma_{\mathrm{hi}})/2$. If $\textsc{CertSign}_j$ returns POS, then $u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p) > 0$. By the definition of $\sigma_j^\star(\omega,p)$, this implies

$$\sigma < \sigma_j^\star(\omega,p),$$

so the update $\sigma_{\mathrm{lo}} \leftarrow \sigma$ preserves the invariant $\sigma_j^\star(\omega,p) \in [\sigma_{\mathrm{lo}}, \sigma_{\mathrm{hi}}]$. If $\textsc{CertSign}_j$ returns NEG, then $u_{\mathrm{NO}}^{(j)}(\sigma;\omega,p) < 0$, hence

$$\sigma \ge \sigma_j^\star(\omega,p),$$

and the update $\sigma_{\mathrm{hi}} \leftarrow \sigma$ also preserves the invariant.

Each bisection step halves the interval length. Therefore after m bisection updates,

$$\sigma_{\mathrm{hi}} - \sigma_{\mathrm{lo}} \le 2^{-m}\big(\sigma_{\mathrm{hi}}^{\mathrm{init}} - \sigma_{\mathrm{lo}}^{\mathrm{init}}\big).$$

Thus, once

$$m \ge \left\lceil \log_2\left(\frac{\sigma_{\mathrm{hi}}^{\mathrm{init}} - \sigma_{\mathrm{lo}}^{\mathrm{init}}}{\epsilon}\right)\right\rceil,$$

we have $\sigma_{\mathrm{hi}} - \sigma_{\mathrm{lo}} \le \epsilon$. Since $\sigma_j^\star(\omega,p) \in [\sigma_{\mathrm{lo}}, \sigma_{\mathrm{hi}}]$, the returned midpoint satisfies

$$|\hat{\sigma} - \sigma_j^\star(\omega,p)| \le \frac{\sigma_{\mathrm{hi}} - \sigma_{\mathrm{lo}}}{2} \le \epsilon.$$

□

E.3. Witness extraction

For the exchange layer, localizing the scalar value $\sigma_j^\star(\omega,p)$ is not by itself enough. The mirror update requires a cut vector, which is obtained from a near-minimizing sign-flipped kernel. We therefore impose the following deterministic witness contract. It is the same policywise optimization problem as in Appendix D, but the routine must additionally return a primal witness with a certified objective guarantee.

Definition E.2 (Certified policywise witness contract)**.** Fix $j \in [J]$, $p \in \mathcal{P}_{\mathrm{test}}^{1:J}$, an allocation $\omega \in \Delta(\mathcal{S} \times \mathcal{A})$ with strictly positive coordinates, and an accuracy $\epsilon > 0$. A certified policywise witness routine returns a kernel $q_j^\epsilon \in \mathrm{Alt}_j^{\lessgtr}(p)$ and a scalar $\widehat{\sigma}_j$ such that

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

where

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

Equivalently, the returned witness satisfies

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

Remark E.3 (Implementation of the witness contract). The optimization variable in the policywise frontier problem is the transition kernel q, equivalently a product of row variables. In implementations based on the reversed-MDP or product-budget formulation, feasible primal iterates are therefore kernels. A routine satisfies Definition E.2 if, in addition to certifying the scalar frontier value, it keeps a feasible opposite-sign primal kernel q_j^ϵ whose objective value is certified to be within the requested tolerance of the policywise frontier. The exchange analysis uses only this deterministic witness certificate; it does not require the static-sampling statistical guarantee of Ariu et al. (2025), nor does it rely on any uncertified primal iterate.

Algorithm 3 Certified bottleneck cut oracle

Require: model $p' \in \mathcal{P}_{\text{test}}^{1:J}$, query point $x \in K_\mu$, accuracy $\xi^{\text{cut}} > 0$
1: **for** $j = 1, \dots, J$ **do**
2: Run the certified policywise witness routine with accuracy $\xi^{\text{cut}}/4$, model p', allocation x, and target policy j
3: Obtain $q_j \in \text{Alt}_j^{\leq}(p')$ and $\widehat{\sigma}_j$
4: **end for**
5: Choose $j_\star \in \arg\min_{j \in [J]} \widehat{\sigma}_j$
6: Set $q_\star \leftarrow q_{j_\star}$
7: Set

$$g_{sa} \leftarrow \text{KL}\big(p'(\cdot \mid s, a) \,\|\, q_\star(\cdot \mid s, a)\big), \qquad (s, a) \in \mathcal{S} \times \mathcal{A}$$

8: **return** $(j_\star, q_\star, g)$

E.4. Instantiation of the bottleneck cut oracle

We now show that the certified policywise witness contract gives the cut oracle used by PT-ACE(μ). For $p' \in \mathcal{P}_{\text{test}}^{1:J}$, define the closed policywise and simultaneous frontiers

$$F_{p'}^{(j)}(x) := \inf_{q \in \text{Alt}_j^{\leq}(p')} \Phi(p'; x, q), \qquad F_{p'}^{1:J}(x) := \min_{j \in [J]} F_{p'}^{(j)}(x).$$

The exchange layer needs an inexact supergradient cut for the concave function $x \mapsto F_{p'}^{1:J}(x)$, which is a pointwise infimum of linear functions.

Proposition E.4 (Certified bottleneck cut)**.** *Fix $p' \in \mathcal{P}_{\text{test}}^{1:J}$, $x \in K_\mu$, and $\xi^{\text{cut}} > 0$. Suppose Algorithm 3 is run with policywise witness accuracy $\xi^{\text{cut}}/4$. Let $(j_\star, q_\star, g)$ be its output. Then*

$$q_\star \in \text{Alt}_{j_\star}^{\leq}(p')$$

and

$$g_{sa} = \text{KL}\big(p'(\cdot \mid s, a) \,\|\, q_\star(\cdot \mid s, a)\big).$$

Moreover, for every $z \in \Delta(\mathcal{S} \times \mathcal{A})$,

$$F_{p'}^{1:J}(z) - F_{p'}^{1:J}(x) \leq \langle g, z - x \rangle + \xi^{\text{cut}}. \tag{8}$$

In particular, (8) holds for every $z \in \widetilde{\Sigma}_{p'}^{\mu}$.

Proof. For each j, the witness contract gives

$$F_{p'}^{(j)}(x) \leq \widehat{\sigma}_j \leq F_{p'}^{(j)}(x) + \frac{\xi^{\text{cut}}}{4},$$

and

$$\Phi(p'; x, q_j) \leq \widehat{\sigma}_j + \frac{\xi^{\text{cut}}}{4} \leq F_{p'}^{(j)}(x) + \frac{\xi^{\text{cut}}}{2}.$$

Let $j^\circ \in \arg\min_{j \in [J]} F_{p'}^{(j)}(x)$, so that $F_{p'}^{1:J}(x) = F_{p'}^{(j^\circ)}(x)$. Since $j_\star$ minimizes $\widehat{\sigma}_j$,

$$\widehat{\sigma}_{j_\star} \leq \widehat{\sigma}_{j^\circ} \leq F_{p'}^{(j^\circ)}(x) + \frac{\xi^{\text{cut}}}{4} = F_{p'}^{1:J}(x) + \frac{\xi^{\text{cut}}}{4}.$$

Using the witness contract for $j_\star$,

$$\Phi(p'; x, q_\star) \leq \widehat{\sigma}_{j_\star} + \frac{\xi^{\text{cut}}}{4} \leq F_{p'}^{1:J}(x) + \frac{\xi^{\text{cut}}}{2}.$$

Increasing the constant from $\xi^{\text{cut}}/2$ to ξ^{cut} leaves room for numerical tolerances in the witness routine.

Now fix any $z \in \Delta(\mathcal{S} \times \mathcal{A})$. Since $q_\star \in \mathrm{Alt}^{\leq}_{j_\star}(p')$, we have

$$F^{1:J}_{p'}(z) \leq F^{(j_\star)}_{p'}(z) \leq \Phi(p'; z, q_\star).$$

By the definition of g,

$$\Phi(p'; z, q_\star) = \Phi(p'; x, q_\star) + \langle g, z - x \rangle.$$

Combining the last three displays gives

$$F^{1:J}_{p'}(z) \leq F^{1:J}_{p'}(x) + \langle g, z - x \rangle + \xi^{\mathrm{cut}},$$

which is exactly (8). □

Computational remarks. The construction above reduces each active cut-oracle call to J independent certified policywise witness computations. Each computation has the same fixed-policy frontier objective and product-kernel feasible set as in Appendix D, but it must return the additional primal witness specified in Definition E.2. Selecting the smallest certified frontier value produces the bottleneck index, and the cut vector is obtained by evaluating the corresponding rowwise KL divergences. The remaining exchange update consists of two convex KL projections over fixed-floor occupancy polytopes. The theoretical analysis only requires that the prescribed cut tolerances ξ^{cut}_t are met and vanish along the run.

Remark E.5 (Use inside PT-ACE(μ)). In Algorithm 1, the call

$$(j_t, q_t, g_t) \leftarrow \mathrm{CutOracle}(\bar{p}_t, x_t, \xi^{\mathrm{cut}}_t)$$

is implemented by Algorithm 3 with $(p', x, \xi^{\mathrm{cut}}) = (\bar{p}_t, x_t, \xi^{\mathrm{cut}}_t)$. Hence the oracle contract used in the exchange analysis,

$$F^{1:J}_{\bar{p}_t}(z) - F^{1:J}_{\bar{p}_t}(x_t) \leq \langle g_t, z - x_t \rangle + \xi^{\mathrm{cut}}_t, \qquad z \in \widetilde{\Sigma}^{\mu}_{\bar{p}_t},$$

is a consequence of Proposition E.4.

Role in the paper. The bisection part of this appendix shows how certified scalar sign queries can localize the policywise threshold. The witness-extraction and bottleneck-cut parts state the additional certified primal-witness contract needed by the exchange layer and prove that this contract yields a valid inexact bottleneck cut. Thus the stopping layer uses certified scalar brackets, while the exchange layer uses certified primal witnesses. The main theorem is stated at the level of these two deterministic numerical subroutines.

F. Concentration of the empirical kernels

In this appendix we justify the kernel-concentration event

$$\mathcal{C}_T^{\ker}(\zeta) := \bigcap_{t=\tau_T}^{T} \{\|\hat{p}_t - p\|_\infty \le \zeta\}, \qquad \tau_T := \lceil T^{1/4} \rceil,$$

which is the good event used later in Appendix G. Throughout this appendix, for a kernel-valued estimate $\hat{p}$, we write

$$\|\hat{p} - p\|_\infty := \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\hat{p}(\cdot \mid s,a) - p(\cdot \mid s,a)\|_1.$$

We also write

$$N_{sa}(t) := \sum_{u=0}^{t-1} \mathbf{1}\{(S_u, A_u) = (s,a)\}, \qquad \hat{w}_{sa}(t) := \frac{N_{sa}(t)}{t \vee 1},$$

and $\hat{w}(t) := (\hat{w}_{sa}(t))_{(s,a)\in\mathcal{S}\times\mathcal{A}}$.

The proof follows two steps. First, the persistent action floor of PT-ACE(μ) and the anchor-policy ergodicity assumption imply that every state-action pair is visited linearly often with high probability. Second, conditional independence is not invoked through the random count $N_{sa}(t)$. Instead, we prove empirical kernel concentration by applying martingale concentration directly to the row-wise centered transition indicators. This avoids the invalid step of conditioning on the adaptive event $N_{sa}(t) = n$.

F.1. Uniform interiority of the frozen policies

Lemma F.1 (Uniform action floor)**.** *For every $t \ge 0$, the frozen navigation policy χ_t governing $Z_t = (S_t, A_t) \to Z_{t+1} = (S_{t+1}, A_{t+1})$ satisfies*

$$\chi_t(a \mid s) \ge \mu, \qquad \forall (s,a) \in \mathcal{S}\times\mathcal{A}.$$

Consequently, if

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

and

$$P_t\big((s,a),(s',a')\big) := p(s' \mid s,a)\chi_t(a' \mid s'),$$

then for all $t \ge 0$ and all $(s,a),(s',a') \in \mathcal{S}\times\mathcal{A}$,

$$P_t\big((s,a),(s',a')\big) \ge |\mathcal{A}|\mu\, P_p^{\mathrm{Unif}_\mathcal{A}}\big((s,a),(s',a')\big).$$

Proof. By construction, PT-ACE(μ) defines the frozen navigation policy χ_t from the averaged allocation $\bar{x}_t \in K_\mu$:

$$\chi_t(a \mid s) = \frac{\bar{x}_t(s,a)}{\sum_{b\in\mathcal{A}} \bar{x}_t(s,b)}.$$

Since $\bar{x}_t \in K_\mu$, we have $\bar{x}_t(s,a) \ge \mu$ for every (s,a), and $\sum_{b\in\mathcal{A}} \bar{x}_t(s,b) \le 1$. Therefore

$$\chi_t(a \mid s) \ge \bar{x}_t(s,a) \ge \mu,$$

which proves the first claim.

For the second claim, using $\mathrm{Unif}_\mathcal{A}(a' \mid s') = 1/|\mathcal{A}|$ and the lower bound just proved,

$$P_t\big((s,a),(s',a')\big) = p(s' \mid s,a)\chi_t(a' \mid s') \ge p(s' \mid s,a)\mu = |\mathcal{A}|\mu\, p(s' \mid s,a)\mathrm{Unif}_\mathcal{A}(a' \mid s'),$$

as required. □

F.2. Common minorization and uniform ergodicity

Lemma F.2 (Common product Doeblin minorization)**.** *Assume Assumption 3.6. Then there exist an integer $m_\mu \geq 1$, a constant $\varepsilon_\mu \in (0,1)$, and a probability distribution ν_μ on $\mathcal{S} \times \mathcal{A}$, depending only on (μ, p), such that for every integer $i \geq 0$,*

$$P_i P_{i+1} \cdots P_{i+m_\mu-1}(z,\cdot) \geq \varepsilon_\mu \nu_\mu(\cdot), \qquad \forall z \in \mathcal{S} \times \mathcal{A}.$$

In particular, for every $t \geq 0$,

$$P_t^{m_\mu}(z,\cdot) \geq \varepsilon_\mu \nu_\mu(\cdot), \qquad \forall z \in \mathcal{S} \times \mathcal{A}.$$

Proof. Let $P_u := P_p^{\mathrm{Unif}_\mathcal{A}}$ be the state-action transition kernel induced by the uniform policy. By Assumption 3.6, the Markov chain on $\mathcal{S} \times \mathcal{A}$ with kernel P_u is irreducible and aperiodic. Since $\mathcal{S} \times \mathcal{A}$ is finite, there exists an integer $m_\mu \geq 1$ such that

$$P_u^{m_\mu}(z,z') > 0, \qquad \forall z, z' \in \mathcal{S} \times \mathcal{A}.$$

Let ω_u denote the stationary distribution of P_u, and define

$$\sigma_u := \min_{z,z' \in \mathcal{S}\times\mathcal{A}} \frac{P_u^{m_\mu}(z,z')}{\omega_u(z')}.$$

Because $P_u^{m_\mu}(z,z') > 0$ for all z, z' and $\omega_u(z') > 0$ for all z', we have $\sigma_u > 0$. Hence

$$P_u^{m_\mu}(z,\cdot) \geq \sigma_u \omega_u(\cdot), \qquad \forall z \in \mathcal{S} \times \mathcal{A}.$$

By Lemma F.1, $P_t \geq |\mathcal{A}|\mu\, P_u$ entrywise for every $t \geq 0$. Fix an integer $i \geq 0$. Since all matrices involved are nonnegative, repeated multiplication preserves entrywise order, and therefore

$$P_i P_{i+1} \cdots P_{i+m_\mu-1} \geq (|\mathcal{A}|\mu)^{m_\mu} P_u^{m_\mu}$$

entrywise. Consequently, for every $z \in \mathcal{S} \times \mathcal{A}$,

$$P_i P_{i+1} \cdots P_{i+m_\mu-1}(z,\cdot) \geq (|\mathcal{A}|\mu)^{m_\mu} P_u^{m_\mu}(z,\cdot) \geq (|\mathcal{A}|\mu)^{m_\mu} \sigma_u \omega_u(\cdot).$$

Thus the conclusion holds with

$$\varepsilon_\mu := (|\mathcal{A}|\mu)^{m_\mu} \sigma_u, \qquad \nu_\mu := \omega_u.$$

The second display follows by taking $i = t$ and observing that $P_t P_t \cdots P_t = P_t^{m_\mu}$. □

F.3. High-probability lower bounds on state-action visit counts

Lemma F.3 (Visit-count lower bound at a fixed time)**.** *There exist constants $c_{\mu,p}, C_{\mu,p} > 0$ such that for every $(s,a) \in \mathcal{S}\times\mathcal{A}$ and every sufficiently large t,*

$$\mathbb{P}(N_{sa}(t) \leq c_{\mu,p} t) \leq \exp(-C_{\mu,p} t).$$

Consequently,

$$\mathbb{P}\Big(\min_{(s,a)\in\mathcal{S}\times\mathcal{A}} N_{sa}(t) \leq c_{\mu,p} t\Big) \leq |\mathcal{S}||\mathcal{A}| \exp(-C_{\mu,p} t).$$

Proof. Let $m_\mu, \varepsilon_\mu, \nu_\mu$ be given by Lemma F.2, and define

$$q_{sa} := \varepsilon_\mu \nu_\mu(s,a), \qquad q_\star := \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} q_{sa}.$$

Since ν_μ has full support on the finite space $\mathcal{S} \times \mathcal{A}$, we have $q_\star > 0$.

Fix $(s,a) \in \mathcal{S} \times \mathcal{A}$. For each integer $k \geq 1$, define the block-endpoint indicator $Y_k^{sa} := \mathbf{1}\{Z_{km_\mu} = (s,a)\}$, and let $\mathcal{G}_k := \sigma(Z_0, \ldots, Z_{km_\mu})$. Then $(\mathcal{G}_k)_{k\geq 0}$ is a filtration, and Y_k^{sa} is $\mathcal{G}_k$-measurable.

By Lemma F.2, for every $k \geq 1$ and every realization of $Z_{(k-1)m_\mu}$,

$$P_{(k-1)m_\mu} P_{(k-1)m_\mu+1} \cdots P_{km_\mu-1}\big(Z_{(k-1)m_\mu}, \cdot\big) \geq \varepsilon_\mu \nu_\mu(\cdot).$$

Therefore,

$$\mathbb{E}[Y_k^{sa} \mid \mathcal{G}_{k-1}] = \mathbb{P}\big(Z_{km_\mu} = (s,a) \mid \mathcal{G}_{k-1}\big) \geq \varepsilon_\mu \nu_\mu(s,a) = q_{sa} \geq q_\star.$$

Hence, for every $n \geq 1$,

$$\sum_{k=1}^{n} \mathbb{E}[Y_k^{sa} \mid \mathcal{G}_{k-1}] \geq n q_\star.$$

Define

$$M_n^{sa} := \sum_{k=1}^{n} \Big(Y_k^{sa} - \mathbb{E}[Y_k^{sa} \mid \mathcal{G}_{k-1}] \Big).$$

Then $(M_n^{sa})_{n\geq 0}$ is a martingale with respect to $(\mathcal{G}_n)_{n\geq 0}$, and $|M_n^{sa} - M_{n-1}^{sa}| \leq 1$. Azuma–Hoeffding gives

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

Let $n_t := \lfloor t/m_\mu \rfloor$. Since each event $\{Z_{km_\mu} = (s,a)\}$ contributes at least one visit to (s,a) by time t, we have $N_{sa}(t) \geq \sum_{k=1}^{n_t} Y_k^{sa}$. Thus

$$\mathbb{P}\Big(N_{sa}(t) \leq \frac{q_\star}{2} n_t \Big) \leq \exp\left(-\frac{q_\star^2}{8} n_t \right).$$

For all sufficiently large t, $n_t \geq t/(2m_\mu)$, and therefore

$$\mathbb{P}\left(N_{sa}(t) \leq \frac{q_\star}{4m_\mu} t \right) \leq \exp\left(-\frac{q_\star^2}{16 m_\mu} t \right).$$

Hence the claim holds with

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

The second display follows by a union bound over $(s,a) \in \mathcal{S} \times \mathcal{A}$. □

Corollary F.4 (Tail-window visit lower bound)**.** *With the constants $c_{\mu,p}, C_{\mu,p}$ from Lemma F.3, for every sufficiently large T,*

$$\mathbb{P}\left(\exists t \in [\tau_T, T] : \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} N_{sa}(t) \leq c_{\mu,p} t \right) \leq |\mathcal{S}||\mathcal{A}|\, T \, \exp(-C_{\mu,p}\tau_T).$$

Proof. By Lemma F.3 and a union bound over $t \in \{\tau_T, \tau_T + 1, \ldots, T\}$,

$$\mathbb{P}\left(\exists t \in [\tau_T, T] : \min_{(s,a)} N_{sa}(t) \leq c_{\mu,p} t \right) \leq \sum_{t=\tau_T}^{T} |\mathcal{S}||\mathcal{A}| \exp(-C_{\mu,p} t).$$

Since the summand is decreasing in t, this is at most

$$|\mathcal{S}||\mathcal{A}|\, T \, \exp(-C_{\mu,p}\tau_T),$$

which proves the claim. □

F.4. Tail-window concentration of the empirical kernels

We now prove concentration of the empirical transition kernels under adaptive sampling. The key point is that we do not condition on the random count $N_{sa}(t)$. Instead, we use martingale concentration for each row and next state.

For notational clarity, in this subsection we write the transition samples as (S_u, A_u, S_{u+1}), $u = 0, \ldots, t-1$, and recall that

$$N_{sa}(t) = \sum_{u=0}^{t-1} \mathbf{1}\{(S_u, A_u) = (s,a)\}.$$

When $N_{sa}(t) \geq 1$, the empirical transition row is

$$\hat{p}_t(y \mid s,a) := \frac{1}{N_{sa}(t)} \sum_{u=0}^{t-1} \mathbf{1}\{(S_u, A_u) = (s,a),\ S_{u+1} = y\}.$$

If $N_{sa}(t) = 0$, the value of $\hat{p}_t(\cdot \mid s,a)$ can be defined arbitrarily; this case is handled by the visit-count lower bound.

Let $\mathcal{F}_u$ denote the history available before the environment draws S_{u+1}. Under the frozen-policy convention, $Z_u = (S_u, A_u)$ and χ_u are $\mathcal{F}_u$-measurable, then

$$S_{u+1} \sim p(\cdot \mid S_u, A_u), \qquad A_{u+1} \sim \chi_u(\cdot \mid S_{u+1}).$$

The empirical transition row $\hat{p}_t(\cdot \mid s,a)$ is updated using the transitions (S_u, A_u, S_{u+1}), $u = 0, \ldots, t-1$.

For each $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $y \in \mathcal{S}$, define

$$M_{sa,y}(t) := \sum_{u=0}^{t-1} \mathbf{1}\{(S_u, A_u) = (s,a)\}\Big(\mathbf{1}\{S_{u+1} = y\} - p(y \mid s,a)\Big).$$

Lemma F.5 (Row-wise transition martingales)**.** *For every $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $y \in \mathcal{S}$, the process $(M_{sa,y}(t))_{t\geq 0}$ is a martingale with respect to $(\mathcal{F}_t)_{t\geq 0}$. Its increments are bounded by 1. Moreover, whenever $N_{sa}(t) \geq 1$,*

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

Proof. For $u \geq 0$, set

$$\Delta_{sa,y}(u) := \mathbf{1}\{(S_u, A_u) = (s,a)\}\Big(\mathbf{1}\{S_{u+1} = y\} - p(y \mid s,a)\Big).$$

Then $M_{sa,y}(t) = \sum_{u=0}^{t-1} \Delta_{sa,y}(u)$.

Since $Z_u = (S_u, A_u)$ is $\mathcal{F}_u$-measurable and, conditional on $\mathcal{F}_u$, the next state has law $p(\cdot \mid S_u, A_u)$, we have

$$\mathbb{E}[\Delta_{sa,y}(u) \mid \mathcal{F}_u] = \mathbf{1}\{(S_u, A_u) = (s,a)\}\Big(p(y \mid S_u, A_u) - p(y \mid s,a)\Big) = 0.$$

Thus $M_{sa,y}$ is a martingale. The increment bound $|\Delta_{sa,y}(u)| \leq 1$ is immediate.

Finally,

$$\sum_{u=0}^{t-1} \mathbf{1}\{(S_u, A_u) = (s,a),\ S_{u+1} = y\} = N_{sa}(t)\hat{p}_t(y \mid s,a),$$

while

$$\sum_{u=0}^{t-1} \mathbf{1}\{(S_u, A_u) = (s,a)\} p(y \mid s,a) = N_{sa}(t) p(y \mid s,a).$$

Subtracting gives

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

and the claimed identity follows whenever $N_{sa}(t) \geq 1$. □

Lemma F.6 (Kernel concentration at a fixed time)**.** *For every $\zeta > 0$, there exist constants $B_{\mu,p,\zeta} > 0$ and $C_{\mu,p,\zeta} > 0$ such that, for every sufficiently large t,*

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

Proof. Fix $\zeta > 0$. Let $c_{\mu,p}, C_{\mu,p} > 0$ be the constants from Lemma F.3, so that, for every (s,a) and every sufficiently large t,

$$\mathbb{P}\big(N_{sa}(t) \leq c_{\mu,p} t\big) \leq \exp(-C_{\mu,p} t).$$

Fix a state-action pair (s, a). On the event $\{N_{sa}(t) > c_{\mu,p}t\}$, if

$$\|\hat{p}_t(\cdot \mid s, a) - p(\cdot \mid s, a)\|_1 > \zeta,$$

then there exists $y \in \mathcal{S}$ such that

$$|\hat{p}_t(y \mid s, a) - p(y \mid s, a)| > \frac{\zeta}{|\mathcal{S}|}.$$

By Lemma F.5, this implies

$$|M_{sa,y}(t)| = N_{sa}(t)\, |\hat{p}_t(y \mid s, a) - p(y \mid s, a)| > \frac{c_{\mu,p}\zeta}{|\mathcal{S}|}\, t.$$

For each fixed (s, a, y), Azuma–Hoeffding's inequality applied to the martingale $M_{sa,y}$ gives

$$\mathbb{P}\left(|M_{sa,y}(t)| > \frac{c_{\mu,p}\zeta}{|\mathcal{S}|}\, t\right) \leq 2\exp\left(-\frac{c_{\mu,p}^2\zeta^2}{2|\mathcal{S}|^2}\, t\right).$$

Therefore,

$$\begin{aligned}\mathbb{P}\left(\|\hat{p}_t(\cdot \mid s, a) - p(\cdot \mid s, a)\|_1 > \zeta\right) &\leq \mathbb{P}\big(N_{sa}(t) \leq c_{\mu,p}t\big) \\ &\quad + \sum_{y\in\mathcal{S}} \mathbb{P}\left(|M_{sa,y}(t)| > \frac{c_{\mu,p}\zeta}{|\mathcal{S}|}\, t\right) \\ &\leq \exp(-C_{\mu,p}t) + 2|\mathcal{S}|\exp\left(-\frac{c_{\mu,p}^2\zeta^2}{2|\mathcal{S}|^2}\, t\right).\end{aligned}$$

Taking a union bound over $(s, a) \in \mathcal{S} \times \mathcal{A}$, and using

$$\|\hat{p}_t - p\|_\infty = \max_{(s,a)\in\mathcal{S}\times\mathcal{A}} \|\hat{p}_t(\cdot \mid s, a) - p(\cdot \mid s, a)\|_1,$$

we obtain

$$\mathbb{P}\big(\|\hat{p}_t - p\|_\infty > \zeta\big) \leq |\mathcal{S}||\mathcal{A}| \left[\exp(-C_{\mu,p}t) + 2|\mathcal{S}|\exp\left(-\frac{c_{\mu,p}^2\zeta^2}{2|\mathcal{S}|^2}\, t\right)\right].$$

Thus the claim holds with

$$B_{\mu,p,\zeta} := |\mathcal{S}||\mathcal{A}|(1 + 2|\mathcal{S}|)$$

and

$$C_{\mu,p,\zeta} := \min\left\{C_{\mu,p}, \frac{c_{\mu,p}^2\zeta^2}{2|\mathcal{S}|^2}\right\}.$$

□

Proposition F.7 (Tail-window kernel concentration)**.** *Fix $\zeta > 0$ and define*

$$\mathcal{C}_T^{\ker}(\zeta) := \bigcap_{t=\tau_T}^{T} \{\|\hat{p}_t - p\|_\infty \leq \zeta\}, \qquad \tau_T := \lceil T^{1/4} \rceil.$$

Then there exist constants $B_{\mu,p,\zeta} > 0$ and $C_{\mu,p,\zeta} > 0$ such that, for all sufficiently large T,

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

Proof. By Lemma F.6 and a union bound over $t \in \{\tau_T, \tau_T + 1, \ldots, T\}$,

$$\begin{aligned}\mathbb{P}\big((\mathcal{C}_T^{\ker}(\zeta))^c\big) &= \mathbb{P}\left(\exists t \in \{\tau_T, \ldots, T\} : \|\hat{p}_t - p\|_\infty > \zeta\right) \\ &\leq \sum_{t=\tau_T}^{T} B_{\mu,p,\zeta}\exp(-C_{\mu,p,\zeta}t).\end{aligned}$$

Since the summand is decreasing in t, the right-hand side is at most

$$B_{\mu,p,\zeta}\, T\, \exp(-C_{\mu,p,\zeta}\tau_T),$$

which proves the claim. □

G. Exchange analysis for PT-ACE(μ)

The goal of this appendix is to analyze the exchange layer of PT-ACE(μ). We show that, on the kernel good event $\mathcal{C}_T^{\ker}(\zeta)$, the weighted average of the corrected exchange iterates is close to maximizing the true simultaneous policy-testing objective.

Closed simultaneous-testing frontier used by the exchange layer. For any kernel $p' \in \mathcal{P}$ and any policy index $j \in [J]$, define the policywise closed sign-testing alternative set

$$\mathrm{Alt}_j^{\leq}(p') := \{q \in \mathcal{P} : \ V_{p',j}(\rho)V_{q,j}(\rho) \leq 0\}.$$

For an allocation $w \in \Delta(\mathcal{S} \times \mathcal{A})$, define the policywise frontier

$$F_{p'}^{(j)}(w) := \min_{q \in \mathrm{Alt}_j^{\leq}(p')} \Phi(p'; w, q), \qquad \Phi(p'; w, q) := \sum_{s,a} w_{sa} \mathrm{KL}_{sa}(p', q).$$

The simultaneous exchange objective is the bottleneck frontier

$$F_{p'}^{1:J}(w) := \min_{j \in [J]} F_{p'}^{(j)}(w).$$

When $p' \in \mathcal{P}_{\mathrm{test}}^{1:J}$, Appendix C shows that, for each j, the closed and strict testing frontiers agree at level zero. Hence

$$F_{p'}^{1:J}(w) = \min_{j \in [J]} \inf_{q \in \mathrm{Alt}_j(p')} \Phi(p'; w, q), \qquad \mathrm{Alt}_j(p') := \{q \in \mathcal{P} : \ V_{p',j}(\rho)V_{q,j}(\rho) < 0\}.$$

Thus $F_p^{1:J}$ coincides with the simultaneous lower-bound objective used in Theorem 2.3.

The exchange layer does not optimize $F_p^{1:J}$ over Σ_p directly. Instead, at round t, it optimizes the surrogate objective $F_{\bar{p}_t}^{1:J}$ over the surrogate feasible set

$$\widetilde{\Sigma}_{\bar{p}_t}^{\mu} := \Sigma_{\bar{p}_t} \cap K_\mu, \qquad K_\mu := \{w \in \Delta(\mathcal{S} \times \mathcal{A}) : w_{sa} \geq \mu \ \forall (s,a)\}.$$

Here

$$\bar{p}_t(\cdot \mid s, a) := (1 - \theta_t)\hat{p}_t(\cdot \mid s, a) + \theta_t \mathrm{Unif}_{\mathcal{S}}(\cdot), \qquad \theta_t := (t \vee 1)^{-1/4}.$$

Because the domain moves with the data, a fixed optimizer of the true problem is not directly feasible for every mirror step.

The implementation therefore includes the safeguard described in Algorithm 1. If the current or next fixed-μ domain is empty, the exchange step is skipped and the current center is retained. If the current domain is nonempty but the stored center is not feasible for it, the center is first reanchored to an arbitrary point of the current domain before the cut oracle is queried. These safeguards affect only an initial transient in the asymptotic analysis: on the kernel good event, the fixed-μ domains are nonempty throughout the tail window for all sufficiently large T.

Our remedy is to compare the exchange iterates to a *transported benchmark* generated by a fixed interior navigation policy. Let

$$w^\star \in \arg\max_{w \in \Sigma_p} F_p^{1:J}(w),$$

and let $w^{2\mu} \in \Sigma_p \cap K_{2\mu}$ be the corresponding μ-interior benchmark constructed in Subsection G.2. Define

$$\chi^\mu(a \mid s) := \frac{w_{sa}^{2\mu}}{\sum_{b \in \mathcal{A}} w_{sb}^{2\mu}}.$$

Since $w^{2\mu} \in K_{2\mu}$, every action under χ^μ has a uniform positive floor. For each t, define the transported benchmark

$$w_t^\mu := \alpha_{\bar{p}_t}^{\chi^\mu}.$$

By construction, $w_t^\mu \in \Sigma_{\bar{p}_t}$. The perturbation lemmas proved below show that on $\mathcal{C}_T^{\ker}(\zeta)$, for ζ small enough and T large enough, $w_t^\mu \in K_\mu$, hence

$$w_t^\mu \in \widetilde{\Sigma}_{\bar{p}_t}^{\mu} \qquad \forall \tau_T \leq t \leq T.$$

Thus w_t^μ is a feasible comparator for every surrogate mirror step in the tail window and stays close to $w^{2\mu}$, whose value is within the multiplicative factor $1 + c(\mu, p)$ of the true optimum.

The appendix is organized as follows. Subsection G.1 defines the surrogate empirical objective, the active-round predictor–corrector KL-prox update, and the averaged corrected iterates. Subsection G.2 constructs the μ-interior benchmark and the factor $c(\mu, p)$. Subsection G.3 records the basic KL-prox inequalities used in the tail analysis. Subsection G.4 proves the value- and occupancy-perturbation lemmas. Subsection G.5 is the key continuity subsection: all continuity statements are taken on the compact interior slice K_μ, never on the full simplex. Subsection G.6 lower-bounds the cumulative AdaGrad weights. Subsection G.7 controls the transported-comparator drift. Finally, Subsections G.8 and G.9 derive the tail-window near-optimality and flow-defect conclusions used later in the proof of Theorem 3.7. The navigation appendix uses the frozen-policy convention of Algorithm 1; the present appendix only analyzes the exchange averages $\bar{x}_t$ and is unaffected by the one-step delay except through the downstream use of $\bar{x}_t$.

G.1. Surrogate empirical objective, domain, and update

Define the stabilized surrogate kernel

$$\bar{p}_t(\cdot \mid s, a) := (1 - \theta_t)\hat{p}_t(\cdot \mid s, a) + \theta_t \mathrm{Unif}_{\mathcal{S}}(\cdot), \qquad \theta_t := (t \vee 1)^{-1/4},$$

and the fixed-floor surrogate domain

$$\widetilde{\Sigma}^\mu_{\bar{p}_t} := \Sigma_{\bar{p}_t} \cap K_\mu, \qquad K_\mu := \{w \in \Delta(\mathcal{S} \times \mathcal{A}) : w_{sa} \geq \mu \, \forall (s, a)\}.$$

A round t is called *exchange-active* if both $\widetilde{\Sigma}^\mu_{\bar{p}_t}$ and $\widetilde{\Sigma}^\mu_{\bar{p}_{t+1}}$ are nonempty. If t is not exchange-active, the algorithm skips the exchange step:

$$g_t := 0, \qquad G_{t+1} := G_t, \qquad \eta_t := \frac{1}{\sqrt{G_{t+1}}}, \qquad x_{t+1} := x_t.$$

The running average is still updated with the retained center x_{t+1}.

If t is exchange-active but the stored center does not belong to the current domain, the algorithm first reanchors it by choosing an arbitrary point $x_t^\circ \in \widetilde{\Sigma}^\mu_{\bar{p}_t}$ and replacing x_t by x_t°. Throughout this appendix, on an exchange-active round, x_t denotes this post-reanchoring center, so that

$$x_t \in \widetilde{\Sigma}^\mu_{\bar{p}_t}.$$

On an exchange-active round, the certified bottleneck-cut contract gives the following. The cut oracle, queried at $(\bar{p}_t, x_t, \xi_t^{\mathrm{cut}})$, returns a policy index $j_t \in [J]$ and a kernel $q_t \in \mathrm{Alt}^{\leq}_{j_t}(\bar{p}_t)$ such that

$$F^{(j_t)}_{\bar{p}_t}(x_t) = F^{1:J}_{\bar{p}_t}(x_t), \qquad \Phi(\bar{p}_t; x_t, q_t) \leq F^{(j_t)}_{\bar{p}_t}(x_t) + \xi_t^{\mathrm{cut}}.$$

Define the cut vector

$$g_t := \big(\mathrm{KL}_{sa}(\bar{p}_t, q_t)\big)_{s,a}, \qquad \Phi(\bar{p}_t; w, q_t) = \langle g_t, w \rangle.$$

The AdaGrad accumulator and stepsize are

$$G_{t+1} := G_t + \|g_t\|_\infty^2, \qquad \eta_t := \frac{1}{\sqrt{G_{t+1}}}, \qquad G_0 = \varepsilon_0^2 > 0.$$

The active exchange update has two stages.

Predictor step. After receiving the cut g_t, compute

$$y_{t+1} \in \arg\max_{x \in \widetilde{\Sigma}^\mu_{\bar{p}_t}} \left\{ \langle g_t, x \rangle - \frac{1}{\eta_t} \mathrm{KL}(x \| x_t) \right\}.$$

Correction step. After the transition (S_t, A_t, S_{t+1}) is collected and the raw empirical kernel is updated from $\hat{p}_t$ to $\hat{p}_{t+1}$, refresh the surrogate kernel from $\bar{p}_t$ to $\bar{p}_{t+1}$, and project the predictor onto the new fixed-floor surrogate domain:

$$x_{t+1} \in \arg\min_{x \in \widetilde{\Sigma}^{\mu}_{\bar{p}_{t+1}}} \mathrm{KL}(x \| y_{t+1}).$$

Thus, on every exchange-active round, the point queried by the cut oracle lies in $\widetilde{\Sigma}^{\mu}_{\bar{p}_t}$, while the corrected next center lies in $\widetilde{\Sigma}^{\mu}_{\bar{p}_{t+1}}$. On every round, active or skipped, the algorithm maintains

$$x_t \in K_\mu, \qquad \bar{x}_t \in K_\mu.$$

Define the weighted average of the corrected or retained centers by

$$\bar{x}_T := \frac{\sum_{t=0}^{T-1} \eta_t x_{t+1}}{\sum_{t=0}^{T-1} \eta_t}.$$

All theorem-facing exchange statements below concern these corrected or retained centers and their weighted averages, not the intermediate predictors.

G.2. μ-interior benchmark and the factor $c(\mu, p)$

Let

$$w^\star \in \arg\max_{w \in \Sigma_p} F_p^{1:J}(w), \qquad \alpha_p := \alpha_p^{\mathrm{Unif}_{\mathcal{A}}} \in \Sigma_p, \qquad \underline{\alpha}_p := \min_{(s,a) \in \mathcal{S} \times \mathcal{A}} \alpha_p(s,a).$$

Lemma G.1 (μ-interior benchmark at level 2μ)**.** *Assume*

$$0 < \mu \le \min\Big\{ \frac{1}{2|\mathcal{S}||\mathcal{A}|}, \frac{\underline{\alpha}_p}{4} \Big\}.$$

Define

$$\theta_\mu := \frac{2\mu}{\underline{\alpha}_p}, \qquad w^{2\mu} := (1-\theta_\mu) w^\star + \theta_\mu \alpha_p.$$

Then

$$w^{2\mu} \in \Sigma_p \cap K_{2\mu},$$

and

$$F_p^{1:J}(w^{2\mu}) \ge (1-\theta_\mu) F_p^{1:J}(w^\star).$$

Equivalently,

$$F_p^{1:J}(w^{2\mu}) \ge \frac{1}{1 + c(\mu,p)} \max_{w \in \Sigma_p} F_p^{1:J}(w), \qquad c(\mu,p) := \frac{\theta_\mu}{1-\theta_\mu}.$$

Moreover, $c(\mu,p) \to 0$ *as* $\mu \downarrow 0$.

Proof. Since both $w^\star$ and α_p belong to Σ_p, so does $w^{2\mu}$. Also,

$$w^{2\mu}_{sa} = (1-\theta_\mu) w^\star_{sa} + \theta_\mu \alpha_p(s,a) \ge \theta_\mu \underline{\alpha}_p = 2\mu,$$

hence $w^{2\mu} \in K_{2\mu}$.

The map $F_p^{1:J}$ is a pointwise infimum of linear forms in w, since it can be written as

$$F_p^{1:J}(w) = \inf_{q \in \cup_{j=1}^{J} \mathrm{Alt}_j^{\le}(p)} \Phi(p; w, q).$$

Thus $F_p^{1:J}$ is concave. Therefore

$$F_p^{1:J}(w^{2\mu}) \ge (1-\theta_\mu) F_p^{1:J}(w^\star) + \theta_\mu F_p^{1:J}(\alpha_p) \ge (1-\theta_\mu) F_p^{1:J}(w^\star),$$

since $F_p^{1:J}(\alpha_p) \geq 0$. The remaining statements follow from

$$1 - \theta_\mu = \frac{1}{1 + c(\mu, p)}, \qquad c(\mu, p) = \frac{\theta_\mu}{1 - \theta_\mu},$$

and $\theta_\mu \to 0$ as $\mu \downarrow 0$. □

Define the benchmark navigation policy

$$\chi^\mu(a \mid s) := \frac{w_{sa}^{2\mu}}{\sum_{b \in \mathcal{A}} w_{sb}^{2\mu}}.$$

Since $w^{2\mu} \in K_{2\mu}$, we have

$$\chi^\mu(a \mid s) \geq 2\mu \qquad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Lemma G.2 (Eventual nonemptiness of the fixed-floor domains)**.** *Assume Assumption 3.6. Fix*

$$0 < \mu \leq \min\Big\{\frac{1}{2|\mathcal{S}||\mathcal{A}|}, \frac{\underline{\alpha}_p}{4}\Big\}.$$

Then there exists $\zeta_{\mathrm{dom}} > 0$ such that, on $\mathcal{C}_T^{\mathrm{ker}}(\zeta_{\mathrm{dom}})$, for all sufficiently large T and every $t \in [\tau_T, T]$,

$$\widetilde{\Sigma}_{\bar{p}_t}^{\mu} \neq \varnothing.$$

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

Proof. Let $\alpha_{p'}^{\mathrm{Unif}_{\mathcal{A}}}$ denote the stationary state-action distribution of the uniform policy under kernel p'. By Assumption 3.6 and finite-state perturbation theory for irreducible aperiodic Markov chains, the map

$$p' \mapsto \alpha_{p'}^{\mathrm{Unif}_{\mathcal{A}}}$$

is continuous in a neighborhood of p. Since

$$\min_{s,a} \alpha_p^{\mathrm{Unif}_{\mathcal{A}}}(s, a) = \underline{\alpha}_p$$

and $\mu \leq \underline{\alpha}_p/4$, there exists $\zeta_{\mathrm{dom}} > 0$ such that

$$\|p' - p\|_\infty \leq 2\zeta_{\mathrm{dom}} \quad \Longrightarrow \quad \min_{s,a} \alpha_{p'}^{\mathrm{Unif}_{\mathcal{A}}}(s, a) \geq 2\mu.$$

On $\mathcal{C}_T^{\mathrm{ker}}(\zeta_{\mathrm{dom}})$, for $t \in [\tau_T, T]$,

$$\|\bar{p}_t - p\|_\infty \leq \|\hat{p}_t - p\|_\infty + \theta_t \|\mathrm{Unif}_{\mathcal{S}} - \hat{p}_t\|_\infty \leq \zeta_{\mathrm{dom}} + 2\theta_t.$$

Since $\theta_t \leq \tau_T^{-1/4} \to 0$ uniformly over $t \geq \tau_T$, for all sufficiently large T,

$$\|\bar{p}_t - p\|_\infty \leq 2\zeta_{\mathrm{dom}} \qquad \forall t \in [\tau_T, T].$$

Therefore

$$\min_{s,a} \alpha_{\bar{p}_t}^{\mathrm{Unif}_{\mathcal{A}}}(s, a) \geq 2\mu.$$

But $\alpha_{\bar{p}_t}^{\mathrm{Unif}_{\mathcal{A}}} \in \Sigma_{\bar{p}_t}$, and hence

$$\alpha_{\bar{p}_t}^{\mathrm{Unif}_{\mathcal{A}}} \in \Sigma_{\bar{p}_t} \cap K_\mu = \widetilde{\Sigma}_{\bar{p}_t}^{\mu}.$$

Thus $\widetilde{\Sigma}_{\bar{p}_t}^{\mu} \neq \varnothing$ throughout the tail window. □

G.3. Basic exchange inequalities

Lemma G.3 (Approximate supergradient inequality)**.** *Under the certified numerical contracts and the tolerance condition* (4)*, on every exchange-active round t, for every $x \in \widetilde{\Sigma}^{\mu}_{\bar{p}_t}$,*

$$F^{1:J}_{\bar{p}_t}(x) - F^{1:J}_{\bar{p}_t}(x_t) \le \langle g_t, x - x_t \rangle + \xi^{\mathrm{cut}}_t.$$

Proof. By definition of j_t,

$$F^{1:J}_{\bar{p}_t}(x_t) = F^{(j_t)}_{\bar{p}_t}(x_t).$$

Also, by definition of $F^{(j_t)}_{\bar{p}_t}$,

$$F^{1:J}_{\bar{p}_t}(x) \le F^{(j_t)}_{\bar{p}_t}(x) \le \Phi(\bar{p}_t; x, q_t) = \langle g_t, x \rangle.$$

The inexact cut oracle guarantees

$$\langle g_t, x_t \rangle = \Phi(\bar{p}_t; x_t, q_t) \le F^{(j_t)}_{\bar{p}_t}(x_t) + \xi^{\mathrm{cut}}_t = F^{1:J}_{\bar{p}_t}(x_t) + \xi^{\mathrm{cut}}_t.$$

Subtracting the two displays proves the claim. □

Lemma G.4 (Uniform cut bound)**.** *Assume Assumption 2.2. Under the certified numerical contracts and the tolerance condition* (4)*, there exists $L_\mu < \infty$, depending only on μ, the problem primitives, and $\sup_t \xi^{\mathrm{cut}}_t$, such that*

$$\|g_t\|_\infty \le L_\mu \qquad \forall t \ge 0.$$

Proof. If t is not exchange-active, then $g_t = 0$, and the claim is trivial. Now fix an exchange-active round t.

By Assumption 2.2, for each $j \in [J]$ there exist kernels q^+_j and q^-_j such that

$$V_{q^+_j, j}(\rho) > 0, \qquad V_{q^-_j, j}(\rho) < 0.$$

By mixing each row with a sufficiently small amount of the uniform distribution on $\mathcal{S}$, we may assume that all rows of q^+_j and q^-_j have full support, while preserving the displayed strict signs. Since $\mathcal{P}$ is compact and the second argument of each rowwise KL divergence is bounded away from zero for these finitely many kernels, the constant

$$C^\circ := \max_{j \in [J]} \max \left\{ \max_{p' \in \mathcal{P}} \max_{s,a} \mathrm{KL}_{sa}(p', q^+_j),\ \max_{p' \in \mathcal{P}} \max_{s,a} \mathrm{KL}_{sa}(p', q^-_j) \right\}$$

is finite.

Let j_t be the bottleneck index returned by the cut oracle. If $V_{\bar{p}_t, j_t}(\rho) \ge 0$, then $q^-_{j_t} \in \mathrm{Alt}^{\le}_{j_t}(\bar{p}_t)$. If $V_{\bar{p}_t, j_t}(\rho) < 0$, then $q^+_{j_t} \in \mathrm{Alt}^{\le}_{j_t}(\bar{p}_t)$. Hence, in either case, there is a full-support comparator $q^\circ_t \in \mathrm{Alt}^{\le}_{j_t}(\bar{p}_t)$ such that

$$\max_{s,a} \mathrm{KL}_{sa}(\bar{p}_t, q^\circ_t) \le C^\circ.$$

Therefore

$$F^{(j_t)}_{\bar{p}_t}(x_t) \le \Phi(\bar{p}_t; x_t, q^\circ_t) = \sum_{s,a} x_t(s,a) \mathrm{KL}_{sa}(\bar{p}_t, q^\circ_t) \le C^\circ,$$

because $x_t \in \Delta(\mathcal{S} \times \mathcal{A})$.

The cut oracle returns q_t and g_t with

$$g_t(s,a) = \mathrm{KL}_{sa}(\bar{p}_t, q_t),$$

and satisfies

$$\langle g_t, x_t \rangle = \Phi(\bar{p}_t; x_t, q_t) \le F^{(j_t)}_{\bar{p}_t}(x_t) + \xi^{\mathrm{cut}}_t.$$

Combining the previous two displays and using $\xi^{\mathrm{cut}}_t \le \bar{\xi}^{\mathrm{cut}} := \sup_u \xi^{\mathrm{cut}}_u < \infty$, we get

$$\langle g_t, x_t \rangle \le C^\circ + \bar{\xi}^{\mathrm{cut}}.$$

Finally, $x_t \in K_\mu$, so $x_t(s,a) \geq \mu$ for every (s,a). Since $g_t \geq 0$ coordinatewise,

$$\mu g_t(s,a) \leq x_t(s,a) g_t(s,a) \leq \langle g_t, x_t \rangle \leq C^\circ + \bar{\xi}^{\mathrm{cut}}.$$

Thus

$$g_t(s,a) \leq \frac{C^\circ + \bar{\xi}^{\mathrm{cut}}}{\mu} \qquad \forall (s,a),$$

and the claim follows with

$$L_\mu := \frac{C^\circ + \bar{\xi}^{\mathrm{cut}}}{\mu}.$$

□

Lemma G.5 (One-step KL-prox inequality for the predictor)**.** *On every exchange-active round t, for every $x \in \widetilde{\Sigma}^\mu_{\bar{p}_t}$,*

$$F^{1:J}_{\bar{p}_t}(x) - F^{1:J}_{\bar{p}_t}(x_t) \leq \frac{\mathrm{KL}(x\|x_t) - \mathrm{KL}(x\|y_{t+1})}{\eta_t} + \frac{\eta_t}{2}\|g_t\|^2_\infty + \xi^{\mathrm{cut}}_t.$$

Proof. Lemma G.3 reduces the proof to the standard mirror-descent inequality

$$\langle g_t, x - x_t \rangle \leq \frac{\mathrm{KL}(x\|x_t) - \mathrm{KL}(x\|y_{t+1})}{\eta_t} + \frac{\eta_t}{2}\|g_t\|^2_\infty.$$

Since y_{t+1} maximizes

$$z \mapsto \langle g_t, z \rangle - \frac{1}{\eta_t}\mathrm{KL}(z\|x_t)$$

over the convex set $\widetilde{\Sigma}^\mu_{\bar{p}_t}$, the first-order optimality condition and the three-point identity for KL give

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

Then write

$$\eta_t \langle g_t, x - x_t \rangle = \eta_t \langle g_t, x - y_{t+1} \rangle + \eta_t \langle g_t, y_{t+1} - x_t \rangle$$

and combine Pinsker with Hölder:

$$-\mathrm{KL}(y_{t+1}\|x_t) + \eta_t \langle g_t, y_{t+1} - x_t \rangle \leq -\frac{1}{2}\|y_{t+1} - x_t\|^2_1 + \eta_t \|g_t\|_\infty \|y_{t+1} - x_t\|_1 \leq \frac{\eta_t^2}{2}\|g_t\|^2_\infty.$$

Dividing by η_t proves the claim. □

Lemma G.6 (KL projection onto the new surrogate domain)**.** *On every exchange-active round t, for every $x \in \widetilde{\Sigma}^\mu_{\bar{p}_{t+1}}$,*

$$\mathrm{KL}(x\|x_{t+1}) \leq \mathrm{KL}(x\|y_{t+1}).$$

Proof. Since

$$x_{t+1} \in \arg\min_{z \in \widetilde{\Sigma}^\mu_{\bar{p}_{t+1}}} \mathrm{KL}(z\|y_{t+1}),$$

the Pythagorean inequality for KL projection onto a convex set gives the result. □

Lemma G.7 (Moving-domain one-step inequality)**.** *Assume that round t is exchange-active. Let $w_t \in \widetilde{\Sigma}^\mu_{\bar{p}_t}$ and $w_{t+1} \in \widetilde{\Sigma}^\mu_{\bar{p}_{t+1}}$. Then*

$$\eta_t\Big(F^{1:J}_{\bar{p}_t}(w_t) - F^{1:J}_{\bar{p}_t}(x_t)\Big) \leq \mathrm{KL}(w_t\|x_t) - \mathrm{KL}(w_{t+1}\|x_{t+1}) + C^{\mathrm{KL}}_\mu \|w_{t+1} - w_t\|_1 + \frac{\eta_t^2}{2}\|g_t\|^2_\infty + \eta_t \xi^{\mathrm{cut}}_t,$$

where $C^{\mathrm{KL}}_\mu < \infty$ depends only on μ and $|\mathcal{S}||\mathcal{A}|$.

Proof. By Lemma G.5, applied with $x = w_t \in \widetilde{\Sigma}^{\mu}_{\bar{p}_t}$,

$$\eta_t\Big(F^{1:J}_{\bar{p}_t}(w_t) - F^{1:J}_{\bar{p}_t}(x_t)\Big) \le \mathrm{KL}(w_t\|x_t) - \mathrm{KL}(w_t\|y_{t+1}) + \frac{\eta_t^2}{2}\|g_t\|_\infty^2 + \eta_t\xi_t^{\mathrm{cut}}.$$

We compare the predictor term to the corrected center in the next domain. Since $w_t, w_{t+1} \in K_\mu$ and $y_{t+1} \in K_\mu$, the map $u \mapsto \mathrm{KL}(u\|y_{t+1})$ is Lipschitz on K_μ, uniformly over $y_{t+1} \in K_\mu$. Indeed, for $d := |\mathcal{S}||\mathcal{A}|$,

$$|\mathrm{KL}(u\|y) - \mathrm{KL}(v\|y)| \le C_\mu^{\mathrm{KL}}\|u - v\|_1 \qquad \forall u, v, y \in K_\mu,$$

for some finite C_μ^{KL}. Hence

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

Moreover $w_{t+1} \in \widetilde{\Sigma}^{\mu}_{\bar{p}_{t+1}}$, and Lemma G.6 gives

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

Combining the three displays proves the claim. □

Lemma G.8 (AdaGrad summation bound)**.** *For every* $0 \le m \le T-1$,

$$\sum_{t=m}^{T-1} \eta_t^2\|g_t\|_\infty^2 \le \log\frac{G_T}{G_m}.$$

Proof. Use

$$\eta_t^2\|g_t\|_\infty^2 = \frac{G_{t+1} - G_t}{G_{t+1}} \le \log\frac{G_{t+1}}{G_t},$$

and sum from $t = m$ to $T-1$. On skipped rounds $g_t = 0$, and the displayed identity still holds. □

G.4. Technical perturbation lemmas

Lemma G.9 (Continuity of the policy value in the kernel)**.** *For every fixed stationary policy χ, the map*

$$p' \longmapsto V^{\chi}_{p'}(\rho)$$

is continuous on $\mathcal{P}$. Consequently, for every fixed policy index j,

$$(p', q) \longmapsto V_{p',j}(\rho)V_{q,j}(\rho)$$

is continuous on $\mathcal{P} \times \mathcal{P}$.

Proof. Write

$$V^{\chi}_{p'}(\rho) = \rho^\top(I - \gamma P^{\chi}_{p'})^{-1}r^{\chi}, \qquad P^{\chi}_{p'}(s, s') := \sum_a \chi(a \mid s)p'(s' \mid s, a), \qquad r^{\chi}(s) := \sum_a \chi(a \mid s)r(s, a).$$

The map $p' \mapsto P^{\chi}_{p'}$ is affine, and matrix inversion is continuous on invertible matrices. Since $I - \gamma P^{\chi}_{p'}$ is invertible for every $p' \in \mathcal{P}$, the claim follows. □

Lemma G.10 (Occupancy perturbation under a fixed policy)**.** *Fix $\underline{\chi} > 0$. Suppose χ satisfies*

$$\chi(a \mid s) \ge \underline{\chi} \qquad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Under Assumption 3.6, there exist $\zeta_\chi > 0$ and $L_{\mathrm{occ}}(\underline{\chi}, p) < \infty$ such that, for all $p', p'' \in \mathcal{P}$ with

$$\|p' - p\|_\infty \le \zeta_\chi, \qquad \|p'' - p\|_\infty \le \zeta_\chi,$$

one has

$$\|\alpha^{\chi}_{p'} - \alpha^{\chi}_{p''}\|_1 \le L_{\mathrm{occ}}(\underline{\chi}, p)\,\|p' - p''\|_\infty.$$

Proof. Let $P^{\chi}_{p'}$ be the state-action kernel induced by (p', χ). Since $\chi(a \mid s) \geq \underline{\chi}$, the family $\{P^{\chi}_{p'}\}$ dominates a positive multiple of the uniform-policy chain. By Assumption 3.6 and finiteness of $\mathcal{S} \times \mathcal{A}$, this yields a common Doeblin minorization for all p' in a small neighborhood of p. Uniform geometric ergodicity then gives Lipschitz continuity of the invariant distribution via the standard perturbation identity on the zero-mass subspace. This is the finite-state perturbation argument used in Appendix H. □

G.5. Objective perturbation on the compact interior slice

Let $w^{2\mu}$ and χ^{μ} be as in Subsection G.2, and define

$$w^{\mu}_t := \alpha^{\chi^{\mu}}_{\bar{p}_t}.$$

Lemma G.11 (Feasibility and proximity of the transported benchmark)**.** *There exists $\zeta_1 > 0$ such that for every $0 < \zeta \leq \zeta_1$, on $\mathcal{C}^{\ker}_T(\zeta)$, one has for all sufficiently large T and all $\tau_T \leq t \leq T$,*

$$w^{\mu}_t \in \widetilde{\Sigma}^{\mu}_{\bar{p}_t},$$

and

$$\|w^{\mu}_t - w^{2\mu}\|_1 \leq L_{\mathrm{occ}}(2\mu, p)\left(\zeta + \theta_t \|\mathrm{Unif}_{\mathcal{S}} - p\|_{\infty}\right).$$

Moreover, for every $\tau_T + 1 \leq t \leq T$,

$$\|w^{\mu}_t - w^{\mu}_{t-1}\|_1 \leq L_{\mathrm{occ}}(2\mu, p)\, \|\bar{p}_t - \bar{p}_{t-1}\|_{\infty}.$$

Proof. Apply Lemma G.10 to the fixed navigation policy χ^{μ}, which satisfies a uniform lower bound determined by the 2μ-interior benchmark. Since $\alpha^{\chi^{\mu}}_{p} = w^{2\mu}$,

$$\|w^{\mu}_t - w^{2\mu}\|_1 \leq L_{\mathrm{occ}}(2\mu, p)\, \|\bar{p}_t - p\|_{\infty}.$$

On $\mathcal{C}^{\ker}_T(\zeta)$,

$$\|\bar{p}_t - p\|_{\infty} \leq \|\hat{p}_t - p\|_{\infty} + \theta_t \|\mathrm{Unif}_{\mathcal{S}} - p\|_{\infty} \leq \zeta + \theta_t \|\mathrm{Unif}_{\mathcal{S}} - p\|_{\infty},$$

which proves the first bound. Since $w^{2\mu} \in K_{2\mu}$, the preceding display implies $w^{\mu}_t \in K_{\mu}$ for ζ small enough and T large enough. By construction $w^{\mu}_t \in \Sigma_{\bar{p}_t}$, hence $w^{\mu}_t \in \widetilde{\Sigma}^{\mu}_{\bar{p}_t}$.

The bound on $\|w^{\mu}_t - w^{\mu}_{t-1}\|_1$ follows by applying the same perturbation lemma to $(\bar{p}_t, \bar{p}_{t-1})$. □

Compact-interior continuity. All continuity statements below are taken on the compact interior slice K_{μ}. This is essential: KL-based objectives need not be continuous at boundary points of the simplex. We therefore never use a continuity modulus for $F^{1:J}_{p'}$ on the full simplex $\Delta(\mathcal{S} \times \mathcal{A})$.

We use the policywise closed/strict frontier equality from Appendix C: on a positive allocation slice and for $p' \in \mathcal{P}^{1:J}_{\mathrm{test}}$, the closed zero-sign frontier equals the strict sign-flipped frontier. This permits strict support-safe near-minimizers in the upper-semicontinuity argument below.

Lemma G.12 (Objective perturbation on $\mathcal{P}_{\zeta_2}(p) \times K_{\mu}$)**.** *Assume Assumption 2.2 and $p \in \mathcal{P}^{1:J}_{\mathrm{test}}$. Then there exists $\zeta_2 > 0$ such that the map*

$$(p', w) \longmapsto F^{1:J}_{p'}(w)$$

is continuous on

$$\mathcal{P}_{\zeta_2}(p) \times K_{\mu}, \qquad \mathcal{P}_{\zeta_2}(p) := \{p' \in \mathcal{P} : \|p' - p\|_{\infty} \leq \zeta_2\}.$$

Consequently, for every $0 < \zeta \leq \zeta_2$, there exists a deterministic quantity $\varepsilon_F(\mu, \zeta, T) \geq 0$ such that

$$\lim_{\zeta \downarrow 0} \limsup_{T \to \infty} \varepsilon_F(\mu, \zeta, T) = 0,$$

and on $\mathcal{C}^{\ker}_T(\zeta)$, for every $\tau_T \leq t \leq T$,

$$|F^{1:J}_{\bar{p}_t}(w^{\mu}_t) - F^{1:J}_{p}(w^{2\mu})| \leq \varepsilon_F(\mu, \zeta, T),$$

and, whenever $x_t \in K_{\mu}$,

$$|F^{1:J}_{\bar{p}_t}(x_t) - F^{1:J}_{p}(x_t)| \leq \varepsilon_F(\mu, \zeta, T).$$

Proof. Since $p \in \mathcal{P}_{\text{test}}^{1:J}$, for every $j \in [J]$ we have $V_{p,j}(\rho) \neq 0$. By Lemma G.9, after shrinking $\zeta_2 > 0$ if necessary, we may assume that for every $p' \in \mathcal{P}_{\zeta_2}(p)$ and every $j \in [J]$,

$$\operatorname{sign}(V_{p',j}(\rho)) = \operatorname{sign}(V_{p,j}(\rho)).$$

Fix $j \in [J]$, and write

$$b_j := \operatorname{sign}(V_{p,j}(\rho)).$$

On $\mathcal{P}_{\zeta_2}(p)$, the closed opposite-sign set is frozen:

$$\Gamma_j := \{q \in \mathcal{P} : \ b_j V_{q,j}(\rho) \leq 0\}.$$

Thus, for every $p' \in \mathcal{P}_{\zeta_2}(p)$,

$$F_{p'}^{(j)}(w) = \min_{q \in \Gamma_j} \Phi(p'; w, q).$$

The set Γ_j is compact, since it is a closed subset of the compact kernel space $\mathcal{P}$.

We first prove lower semicontinuity of

$$(p', w) \mapsto F_{p'}^{(j)}(w)$$

on $\mathcal{P}_{\zeta_2}(p) \times K_\mu$. Let

$$(p_n, w_n) \to (p_0, w_0)$$

with $p_n, p_0 \in \mathcal{P}_{\zeta_2}(p)$ and $w_n, w_0 \in K_\mu$. For each n, choose $q_n \in \Gamma_j$ attaining the minimum:

$$F_{p_n}^{(j)}(w_n) = \Phi(p_n; w_n, q_n).$$

By compactness of Γ_j, after passing to a subsequence we may assume $q_n \to q_\infty \in \Gamma_j$. Since the rowwise KL divergence is lower semicontinuous, Φ is lower semicontinuous in (p', w, q). Therefore

$$F_{p_0}^{(j)}(w_0) \leq \Phi(p_0; w_0, q_\infty) \leq \liminf_{n \to \infty} \Phi(p_n; w_n, q_n) = \liminf_{n \to \infty} F_{p_n}^{(j)}(w_n).$$

This proves lower semicontinuity.

We now prove upper semicontinuity. Fix

$$(p_n, w_n) \to (p_0, w_0)$$

in $\mathcal{P}_{\zeta_2}(p) \times K_\mu$, and let $\varepsilon > 0$. Because $p_0 \in \mathcal{P}_{\text{test}}^{1:J}$ and $w_0 \in K_\mu$ has strictly positive coordinates, the policywise closed/strict frontier equality from Appendix C gives

$$\min_{q \in \Gamma_j} \Phi(p_0; w_0, q) = \inf_{\substack{q \in \mathcal{P}: \\ b_j V_{q,j}(\rho) < 0}} \Phi(p_0; w_0, q).$$

Consequently, there exists a strict opposite-sign kernel $q_\varepsilon^{\text{str}} \in \mathcal{P}$ such that

$$b_j V_{q_\varepsilon^{\text{str}},j}(\rho) < 0$$

and

$$\Phi(p_0; w_0, q_\varepsilon^{\text{str}}) \leq F_{p_0}^{(j)}(w_0) + \varepsilon.$$

In particular, the left-hand side is finite.

We next replace $q_\varepsilon^{\text{str}}$ by a full-support strict opposite-sign kernel with nearly the same objective value. By Assumption 2.2, there exists a kernel $q_j^\circ \in \mathcal{P}$ with full row support and strict opposite sign:

$$b_j V_{q_j^\circ,j}(\rho) < 0.$$

For $\gamma' \in (0, 1)$, define

$$q_{\varepsilon,\gamma'} := (1 - \gamma') q_\varepsilon^{\text{str}} + \gamma' q_j^\circ.$$

Then $q_{\varepsilon,\gamma'}$ has full row support for every $\gamma' > 0$. Moreover, since $q \mapsto V_{q,j}(\rho)$ is continuous and $q_\varepsilon^{\mathrm{str}}$ has strict opposite sign, there exists $\gamma_1 > 0$ such that for every $0 < \gamma' \le \gamma_1$,

$$b_j V_{q_{\varepsilon,\gamma'},j}(\rho) < 0.$$

Thus $q_{\varepsilon,\gamma'} \in \Gamma_j$ for all sufficiently small $\gamma' > 0$.

We also have

$$\Phi(p_0; w_0, q_{\varepsilon,\gamma'}) \to \Phi(p_0; w_0, q_\varepsilon^{\mathrm{str}}) \qquad \text{as } \gamma' \downarrow 0.$$

Indeed, since $w_0 \in K_\mu$, all coordinates of w_0 are strictly positive. The finiteness of $\Phi(p_0; w_0, q_\varepsilon^{\mathrm{str}})$ therefore implies that, for every row (s, a), the support of $q_\varepsilon^{\mathrm{str}}(\cdot \mid s, a)$ contains the support of $p_0(\cdot \mid s, a)$. Moreover, $q_{\varepsilon,\gamma'} \to q_\varepsilon^{\mathrm{str}}$ rowwise and

$$q_{\varepsilon,\gamma'}(\cdot \mid s, a) \ge (1 - \gamma') q_\varepsilon^{\mathrm{str}}(\cdot \mid s, a).$$

For all sufficiently small γ', say $\gamma' \le 1/2$, this gives a uniform domination on the p_0-support. Hence the rowwise logarithms in the KL terms are dominated by a finite vector, and the rowwise KL terms converge. Since there are finitely many rows, the displayed convergence of Φ follows.

Choose $0 < \gamma_\varepsilon \le \gamma_1$ so small that

$$\Phi(p_0; w_0, q_{\varepsilon,\gamma_\varepsilon}) \le F_{p_0}^{(j)}(w_0) + 2\varepsilon.$$

Set

$$\widetilde{q}_\varepsilon := q_{\varepsilon,\gamma_\varepsilon}.$$

Then $\widetilde{q}_\varepsilon \in \Gamma_j$, it has full row support, and

$$\Phi(p_0; w_0, \widetilde{q}_\varepsilon) \le F_{p_0}^{(j)}(w_0) + 2\varepsilon.$$

Because $\widetilde{q}_\varepsilon$ has full row support, the map

$$(p', w) \longmapsto \Phi(p'; w, \widetilde{q}_\varepsilon)$$

is continuous in a neighborhood of (p_0, w_0). Hence, for all sufficiently large n,

$$\Phi(p_n; w_n, \widetilde{q}_\varepsilon) \le \Phi(p_0; w_0, \widetilde{q}_\varepsilon) + \varepsilon \le F_{p_0}^{(j)}(w_0) + 3\varepsilon.$$

Since $\widetilde{q}_\varepsilon \in \Gamma_j$, it is feasible for the minimum defining $F_{p_n}^{(j)}(w_n)$. Therefore

$$F_{p_n}^{(j)}(w_n) \le \Phi(p_n; w_n, \widetilde{q}_\varepsilon) \le F_{p_0}^{(j)}(w_0) + 3\varepsilon$$

for all sufficiently large n. Taking the limsup and then sending $\varepsilon \downarrow 0$ gives upper semicontinuity.

Thus $F_{p'}^{(j)}(w)$ is continuous on $\mathcal{P}_{\zeta_2}(p) \times K_\mu$. Since $J < \infty$, the pointwise minimum $F_{p'}^{1:J}(w) = \min_{j \in [J]} F_{p'}^{(j)}(w)$ is also continuous on this compact set, hence uniformly continuous.

The claimed perturbation bounds now follow from uniform continuity, the estimate

$$\|\bar{p}_t - p\|_\infty \le \zeta + \theta_t \|\mathrm{Unif}_{\mathcal{S}} - p\|_\infty$$

on $\mathcal{C}_T^{\mathrm{ker}}(\zeta)$, the transported-benchmark bound in Lemma G.11, and $\sup_{t \ge \tau_T} \theta_t \to 0$ as $T \to \infty$. □

G.6. Growth of the AdaGrad weights

Lemma G.13 (Lower bound on cumulative AdaGrad weights)**.** *Let L_μ be chosen so that*

$$L_\mu^2 \ge G_0 \qquad \textit{and} \qquad \|g_t\|_\infty \le L_\mu \quad \forall t \ge 0.$$

Then, for every $T \ge 1$,

$$\sum_{t=\tau_T}^{T-1} \eta_t \ge \frac{2}{L_\mu}\Big(\sqrt{T+1} - \sqrt{\tau_T + 1}\Big),$$

after increasing L_μ by at most a universal constant factor.

Proof. The bound $\|g_t\|_\infty \le L_\mu$ is global in time by Lemma G.4; in particular, the AdaGrad accumulator is controlled also on the prefix $t < \tau_T$. By the choice of L_μ,

$$G_{t+1} = G_0 + \sum_{u=0}^{t} \|g_u\|_\infty^2 \le L_\mu^2 + (t+1)L_\mu^2 \le (t+2)L_\mu^2.$$

On skipped rounds $g_t = 0$, and the same bound remains valid. Hence

$$\eta_t = \frac{1}{\sqrt{G_{t+1}}} \ge \frac{1}{L_\mu\sqrt{t+2}}.$$

Summing from $t = \tau_T$ to $T-1$ and using an integral comparison gives the claim, after absorbing a harmless universal constant into L_μ. □

G.7. Comparator drift and suffix-average bookkeeping

Lemma G.14 (Comparator drift bound)**.** *There exists $C_{\mathrm{drift}}(\mu, p) < \infty$ such that for every $0 < \zeta \le \zeta_1$, on $\mathcal{C}_T^{\ker}(\zeta)$, and for all sufficiently large T,*

$$\sum_{t=\tau_T+1}^{T-1} \|w_t^\mu - w_{t-1}^\mu\|_1 \le C_{\mathrm{drift}}(\mu, p) \log T.$$

Proof. By Lemma G.11,

$$\|w_t^\mu - w_{t-1}^\mu\|_1 \le L_{\mathrm{occ}}(2\mu, p)\|\bar{p}_t - \bar{p}_{t-1}\|_\infty.$$

Since $\bar{p}_t = (1-\theta_t)\hat{p}_t + \theta_t \mathrm{Unif}_{\mathcal{S}}$,

$$\|\bar{p}_t - \bar{p}_{t-1}\|_\infty \le \|\hat{p}_t - \hat{p}_{t-1}\|_\infty + 2|\theta_t - \theta_{t-1}|.$$

If the transition observed between times $t-1$ and t belongs to row (s_{t-1}, a_{t-1}), then only that empirical row changes, and

$$\|\hat{p}_t - \hat{p}_{t-1}\|_\infty \le \frac{2}{N_{s_{t-1}a_{t-1}}(t) \vee 1}.$$

Summing over t, the empirical-row contribution is at most $O(\log T)$ after grouping by state-action pair, while $\sum_t |\theta_t - \theta_{t-1}| \le 1$. This yields the bound. □

Lemma G.15 (Shift between query and corrected suffix averages)**.** *For $m < t$, define*

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

Then

$$\|\bar{z}_{m:t} - \bar{x}_{m:t}\|_1 \le \frac{2\eta_m}{W_{m:t}}.$$

Proof. Since G_t is nondecreasing, $\eta_t = 1/\sqrt{G_{t+1}}$ is nonincreasing. Writing the two weighted sums with a common index,

$$\sum_{u=m}^{t-1} \eta_u x_{u+1} = \sum_{v=m+1}^{t} \eta_{v-1} x_v, \qquad \sum_{u=m}^{t-1} \eta_u x_u = \eta_m x_m + \sum_{v=m+1}^{t-1} \eta_v x_v.$$

Therefore

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

because all x_v lie in the simplex. Dividing by $W_{m:t}$ proves the claim. □

Reanchoring on the active tail. On $\mathcal{C}_T^{\mathrm{ker}}(\zeta)$, for all sufficiently large T, the fixed-floor domains $\widetilde{\Sigma}_{\bar{p}_t}^{\mu}$ are nonempty throughout the tail window. Thus, after the first active tail round, the correction step ensures $x_{t+1} \in \widetilde{\Sigma}_{\bar{p}_{t+1}}^{\mu}$, so the next query center is already feasible. Reanchoring can therefore occur at most once at the entrance to the active tail. This single entrance correction contributes only a bounded initial KL term in the telescoping argument below.

G.8. Tail-window near-optimality of the exchange averages

Proposition G.16 (Tail-window near-optimality of the exchange averages)**.** *Assume Assumption 2.2, and let $p \in \mathcal{P}_{\mathrm{test}}^{1:J}$. Suppose that the cut routine satisfies the certified bottleneck-cut contract with tolerances obeying (4). Fix an admissible $\mu > 0$ and $\zeta > 0$. Let*

$$\tau_T := \lceil T^{1/4} \rceil, \qquad \vartheta_T := \lceil T^{3/4} \rceil.$$

Then there exists a deterministic remainder $r_{\mathrm{ex}}^{\mathrm{tail}}(T, \mu, \zeta) \geq 0$ such that on $\mathcal{C}_T^{\mathrm{ker}}(\zeta)$, for all sufficiently large T,

$$\inf_{t \in [\vartheta_T, T]} F_p^{1:J}(\bar{x}_t) \geq \frac{1}{1 + c(\mu, p)} \max_{w \in \Sigma_p} F_p^{1:J}(w) - r_{\mathrm{ex}}^{\mathrm{tail}}(T, \mu, \zeta),$$

and

$$\lim_{\zeta \downarrow 0} \limsup_{T \to \infty} r_{\mathrm{ex}}^{\mathrm{tail}}(T, \mu, \zeta) = 0.$$

Proof. By Lemma G.2, after shrinking ζ if necessary, on $\mathcal{C}_T^{\mathrm{ker}}(\zeta)$ and for all sufficiently large T, every round $u \in [\tau_T, T-1]$ is exchange-active. Hence Lemmas G.3, G.5, and G.6 apply throughout the tail window.

Fix $t \in [\vartheta_T, T]$, and define

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

Here $\bar{z}_{m:t}$ is the suffix average of the query centers, while $\bar{x}_{m:t}$ is the suffix average of the corrected centers.

Apply Lemma G.5 with the feasible comparator $w_u^{\mu} \in \widetilde{\Sigma}_{\bar{p}_u}^{\mu}$, multiply by η_u, and sum from $u = \tau_T$ to $t-1$. Use Lemma G.6 to pass from the predictor to the corrected iterate, and then bound the induced comparator-drift term by Lemma G.14. The resulting inequality is

$$\sum_{u=\tau_T}^{t-1} \eta_u \big(F_{\bar{p}_u}^{1:J}(w_u^{\mu}) - F_{\bar{p}_u}^{1:J}(x_u) \big) \leq C_1(\mu, p) \log T + \frac{1}{2} \log \frac{G_t}{G_{\tau_T}} + \sum_{u=\tau_T}^{t-1} \eta_u \xi_u^{\mathrm{cut}},$$

for a deterministic constant $C_1(\mu, p) < \infty$.

By Lemma G.12,

$$F_{\bar{p}_u}^{1:J}(w_u^{\mu}) \geq F_p^{1:J}(w^{2\mu}) - \varepsilon_F(\mu, \zeta, T),$$

and, since $x_u \in K_\mu$,

$$F_{\bar{p}_u}^{1:J}(x_u) \leq F_p^{1:J}(x_u) + \varepsilon_F(\mu, \zeta, T).$$

Therefore

$$\sum_{u=\tau_T}^{t-1} \eta_u \big(F_p^{1:J}(w^{2\mu}) - F_p^{1:J}(x_u) \big) \leq C_1(\mu, p) \log T + \frac{1}{2} \log \frac{G_t}{G_{\tau_T}} + \sum_{u=\tau_T}^{t-1} \eta_u \xi_u^{\mathrm{cut}} + 2 W_{\tau_T:t} \varepsilon_F(\mu, \zeta, T).$$

By concavity of $F_p^{1:J}$,

$$F_p^{1:J}(\bar{z}_{\tau_T:t}) \geq \frac{1}{W_{\tau_T:t}} \sum_{u=\tau_T}^{t-1} \eta_u F_p^{1:J}(x_u).$$

Hence

$$F_p^{1:J}(\bar{z}_{\tau_T:t}) \geq F_p^{1:J}(w^{2\mu}) - \frac{C_1(\mu, p) \log T + \frac{1}{2} \log \frac{G_t}{G_{\tau_T}} + \sum_{u=\tau_T}^{t-1} \eta_u \xi_u^{\mathrm{cut}}}{W_{\tau_T:t}} - 2\varepsilon_F(\mu, \zeta, T).$$

By Lemma G.13 and $t \geq \vartheta_T$,

$$\inf_{t\in[\vartheta_T,T]} W_{\tau_T:t} \geq \frac{2}{L_\mu}\Big(\sqrt{\vartheta_T} - \sqrt{\tau_T+1}\Big) \longrightarrow \infty.$$

Moreover, Lemma G.4 gives $G_t \leq G_0 + tL_\mu^2$. Therefore the logarithmic terms divided by $W_{\tau_T:t}$ vanish uniformly over $t \in [\vartheta_T, T]$. Finally,

$$\frac{\sum_{u=\tau_T}^{t-1} \eta_u \xi_u^{\mathrm{cut}}}{W_{\tau_T:t}} \leq \sup_{u\geq\tau_T} \xi_u^{\mathrm{cut}},$$

which vanishes as $T \to \infty$ by the tolerance condition (4). Thus the ratio term above vanishes uniformly over $t \in [\vartheta_T, T]$.

By Lemma G.1,

$$F_p^{1:J}(w^{2\mu}) \geq \frac{1}{1+c(\mu,p)} \max_{w\in\Sigma_p} F_p^{1:J}(w).$$

Therefore the desired lower bound holds for the query suffix averages $\bar{z}_{\tau_T:t}$, up to a remainder whose

$$\lim_{\zeta\downarrow 0} \limsup_{T\to\infty}$$

is zero.

We now pass from query suffix averages to corrected suffix averages. By Lemma G.15,

$$\|\bar{z}_{\tau_T:t} - \bar{x}_{\tau_T:t}\|_1 \leq \frac{2\eta_{\tau_T}}{W_{\tau_T:t}} = o_T(1)$$

uniformly over $t \in [\vartheta_T, T]$. Since both averages lie in K_μ, the continuity modulus of $F_p^{1:J}$ on K_μ shows that replacing $\bar{z}_{\tau_T:t}$ by $\bar{x}_{\tau_T:t}$ contributes only an additional $o_T(1)$ term.

It remains to pass from the corrected suffix averages $\bar{x}_{\tau_T:t}$ to the running averages $\bar{x}_t$. Let

$$A_{0:\tau_T} := \sum_{u=0}^{\tau_T-1} \eta_u, \qquad W_{\tau_T:t} := \sum_{u=\tau_T}^{t-1} \eta_u.$$

Since both the prefix average and $\bar{x}_{\tau_T:t}$ belong to the simplex,

$$\|\bar{x}_t - \bar{x}_{\tau_T:t}\|_1 \leq 2\,\frac{A_{0:\tau_T}}{A_{0:\tau_T} + W_{\tau_T:t}} \leq 2\,\frac{A_{0:\tau_T}}{W_{\tau_T:t}}.$$

Moreover,

$$A_{0:\tau_T} \leq \frac{\tau_T}{\sqrt{G_0}},$$

while Lemma G.13 gives, uniformly over $t \in [\vartheta_T, T]$,

$$W_{\tau_T:t} \geq \frac{2}{L_\mu}\Big(\sqrt{\vartheta_T} - \sqrt{\tau_T+1}\Big).$$

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

$$\sup_{t\in[\vartheta_T,T]} \frac{A_{0:\tau_T}}{W_{\tau_T:t}} = O\left(\frac{T^{1/4}}{T^{3/8}}\right) = o(1).$$

Thus

$$\sup_{t\in[\vartheta_T,T]} \|\bar{x}_t - \bar{x}_{\tau_T:t}\|_1 = o_T(1).$$

Again using the continuity modulus of $F_p^{1:J}$ on K_μ, this passage contributes only an additional $o_T(1)$ term. Absorbing all vanishing terms into $r_{\mathrm{ex}}^{\mathrm{tail}}(T, \mu, \zeta)$ completes the proof. □

G.9. Tail-window flow-defect bound

Proposition G.17 (Tail-window flow defect)**.** *Fix an admissible $\mu > 0$ and $\zeta > 0$. Let*

$$\tau_T := \lceil T^{1/4} \rceil, \qquad \vartheta_T := \lceil T^{3/4} \rceil.$$

Then there exists a deterministic remainder $r_{\text{flow}}(T, \zeta) \geq 0$ such that on $\mathcal{C}_T^{\text{ker}}(\zeta)$, for all sufficiently large T,

$$\sup_{t \in [\vartheta_T, T]} \sum_{s \in \mathcal{S}} \left| \sum_{a \in \mathcal{A}} \bar{x}_t(s, a) - \sum_{y \in \mathcal{S}, b \in \mathcal{A}} \bar{x}_t(y, b) p(s \mid y, b) \right| \leq r_{\text{flow}}(T, \zeta),$$

and

$$\lim_{\zeta \downarrow 0} \limsup_{T \to \infty} r_{\text{flow}}(T, \zeta) = 0.$$

Proof. By Lemma G.2, after shrinking ζ if necessary, on $\mathcal{C}_T^{\text{ker}}(\zeta)$ and for all sufficiently large T, every round $u \in [\tau_T, T-1]$ is exchange-active. Hence

$$x_{u+1} \in \widetilde{\Sigma}^{\mu}_{\bar{p}_{u+1}} \subseteq \Sigma_{\bar{p}_{u+1}} \qquad \forall u \in [\tau_T, T-1].$$

For each corrected iterate $x_{u+1} \in \Sigma_{\bar{p}_{u+1}}$, we have exact $\bar{p}_{u+1}$-flow balance:

$$\sum_a x_{u+1}(s, a) = \sum_{y,b} x_{u+1}(y, b) \bar{p}_{u+1}(s \mid y, b).$$

Subtracting the corresponding p-flow gives

$$\sum_s \left| \sum_a x_{u+1}(s, a) - \sum_{y,b} x_{u+1}(y, b) p(s \mid y, b) \right| \leq \|\bar{p}_{u+1} - p\|_\infty.$$

Averaging with weights η_u and using convexity of the ℓ_1 norm,

$$\sum_s \left| \sum_a \bar{x}_{\tau_T:t}(s, a) - \sum_{y,b} \bar{x}_{\tau_T:t}(y, b) p(s \mid y, b) \right| \leq \sup_{v \in [\tau_T+1, t]} \|\bar{p}_v - p\|_\infty.$$

On $\mathcal{C}_T^{\text{ker}}(\zeta)$,

$$\|\bar{p}_v - p\|_\infty \leq \zeta + \theta_v \|\text{Unif}_{\mathcal{S}} - p\|_\infty \leq \zeta + \tau_T^{-1/4} \|\text{Unif}_{\mathcal{S}} - p\|_\infty.$$

Thus the corrected suffix averages have vanishing p-flow defect. Passing from $\bar{x}_{\tau_T:t}$ to $\bar{x}_t$ costs only the same vanishing prefix-weight term as in Proposition G.16. Absorb these contributions into $r_{\text{flow}}(T, \zeta)$. This gives

$$\lim_{\zeta \downarrow 0} \limsup_{T \to \infty} r_{\text{flow}}(T, \zeta) = 0.$$

□

H. Navigation

In this appendix we analyze the navigation layer of PT-ACE(μ). Appendix F already provides the kernel good event

$$\mathcal{C}_T^{\mathrm{ker}}(\zeta) := \bigcap_{t=\tau_T}^{T} \{\|\hat{p}_t - p\|_\infty \le \zeta\}, \qquad \tau_T := \lceil T^{1/4} \rceil,$$

together with a quantitative upper bound on its complement. Our goal here is different: we show that the realized empirical allocation $\hat{w}(T)$ tracks the navigation target generated by the frozen navigation policies, and that this is sufficient to transfer the exchange guarantees of Appendix G to the actual trajectory produced by the algorithm.

The argument has three steps. First, we establish quantitative tracking for the non-homogeneous state-action chain generated by the frozen navigation policies (χ_t), and define a navigation good event $\mathcal{C}_T^{\mathrm{nav}}(\zeta)$ using a tail Cesàro average of the stationary occupancies over the shorter window $[\vartheta_T, T]$, where

$$\vartheta_T := \lceil T^{3/4} \rceil.$$

Second, we use an approximate-flow-to-occupancy lemma to compare the stationary occupancies $\alpha_p^{\chi_t}$ of the frozen navigation policies with the exchange allocations $\bar{x}_t$ from Appendix G. Third, combining this bridge with the tail-window conclusions of Appendix G, we prove that on $\mathcal{C}_T^{\mathrm{ker}}(\zeta) \cap \mathcal{C}_T^{\mathrm{nav}}(\zeta)$, the empirical allocation $\hat{w}(T)$ satisfies

$$F_p^{1:J}(\hat{w}(T)) \ge \frac{1}{1 + c(\mu, p)} \max_{w \in \Sigma_p} F_p^{1:J}(w) - o_{T,\zeta}(1).$$

This is the navigation-side input needed later in the stopping-time upper-bound proof.

Throughout this appendix, χ_t denotes the frozen navigation policy used for the transition

$$Z_t = (S_t, A_t) \longrightarrow Z_{t+1} = (S_{t+1}, A_{t+1}),$$

whereas $\pi^{(j)}$ denotes the j-th fixed target policy in the testing family. Under the one-step delayed convention of Algorithm 1, χ_t is fixed before S_{t+1} is observed and is induced by the current averaged allocation:

$$\chi_t(a \mid s) := \frac{\bar{x}_t(s, a)}{\sum_{b \in \mathcal{A}} \bar{x}_t(s, b)}.$$

After observing S_{t+1}, the next action A_{t+1} is sampled from $\chi_t(\cdot \mid S_{t+1})$. The empirical update and exchange correction then define $\bar{x}_{t+1}$, hence the next frozen policy χ_{t+1}.

For the frozen navigation policies (χ_t), define the tail Cesàro stationary target

$$\widetilde{\alpha}_T^{\vartheta} := \frac{1}{T - \vartheta_T + 1} \sum_{t=\vartheta_T}^{T} \alpha_p^{\chi_t},$$

and the navigation good event

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

H.1. Slow variation of the frozen kernels

Recall that the frozen navigation policy χ_t governing $Z_t \to Z_{t+1}$ is induced by the averaged exchange allocation $\bar{x}_t$:

$$\chi_t(a \mid s) := \frac{\bar{x}_t(s, a)}{\sum_{b \in \mathcal{A}} \bar{x}_t(s, b)}.$$

Let

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

For each $t \geq 0$, let

$$P_t\big((s,a),(s',a')\big) := p(s' \mid s,a)\chi_t(a' \mid s')$$

denote the predictable state-action transition kernel induced by p and the frozen policy χ_t. Equivalently,

$$\mathbb{P}(Z_{t+1} = (s',a') \mid \mathcal{F}_t, Z_t = (s,a)) = P_t\big((s,a),(s',a')\big),$$

where $\mathcal{F}_t$ is the history before drawing S_{t+1}.

Lemma H.1 (Increment of the averaged allocation)**.** *For every $t \geq 1$,*

$$\|\bar{x}_{t+1} - \bar{x}_t\|_1 \leq 2\frac{\eta_t}{A_{t+1}}.$$

Proof. By definition,

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

Hence

$$\bar{x}_{t+1} - \bar{x}_t = \frac{\eta_t}{A_{t+1}}(x_{t+1} - \bar{x}_t).$$

Since $x_{t+1}, \bar{x}_t \in \Delta_{\mathcal{S}\times\mathcal{A}}$,

$$\|\bar{x}_{t+1} - \bar{x}_t\|_1 \leq \frac{\eta_t}{A_{t+1}}\|x_{t+1} - \bar{x}_t\|_1 \leq 2\frac{\eta_t}{A_{t+1}}.$$

□

For $x \in K_\mu$, define

$$x_s := \sum_{a\in\mathcal{A}} x_{sa}, \qquad \chi_x(a \mid s) := \frac{x_{sa}}{x_s}.$$

Lemma H.2 (Lipschitz continuity of the policy map on K_μ)**.** *There exists a constant $L_{\chi,\mu} < \infty$, depending only on μ and $(\mathcal{S},\mathcal{A})$, such that for every $x, y \in K_\mu$,*

$$\max_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} |\chi_x(a \mid s) - \chi_y(a \mid s)| \leq L_{\chi,\mu}\|x - y\|_1.$$

Consequently,

$$\|P_x - P_y\|_{1\to 1} \leq L_{\chi,\mu}\|x - y\|_1,$$

where P_x and P_y are the state-action kernels induced by (p,χ_x) and (p,χ_y), respectively.

Proof. Fix $s \in \mathcal{S}$, and write

$$x_s := \sum_{b\in\mathcal{A}} x_{sb}, \qquad y_s := \sum_{b\in\mathcal{A}} y_{sb}.$$

Since $x, y \in K_\mu$, we have $x_s \geq |\mathcal{A}|\mu$ and $y_s \geq |\mathcal{A}|\mu$. For every $a \in \mathcal{A}$,

$$\left|\frac{x_{sa}}{x_s} - \frac{y_{sa}}{y_s}\right| \leq \frac{|x_{sa} - y_{sa}|}{x_s} + y_{sa}\left|\frac{1}{x_s} - \frac{1}{y_s}\right| \leq \frac{|x_{sa} - y_{sa}|}{|\mathcal{A}|\mu} + \frac{|x_s - y_s|}{(|\mathcal{A}|\mu)^2}.$$

Summing over $a \in \mathcal{A}$ and using

$$|x_s - y_s| \leq \sum_{a\in\mathcal{A}} |x_{sa} - y_{sa}|,$$

we obtain

$$\sum_{a\in\mathcal{A}} |\chi_x(a \mid s) - \chi_y(a \mid s)| \leq \left(\frac{1}{|\mathcal{A}|\mu} + \frac{1}{|\mathcal{A}|\mu^2}\right)\sum_{a\in\mathcal{A}} |x_{sa} - y_{sa}|.$$

Taking the maximum over s and then bounding by $\|x - y\|_1$ proves the first claim.

For the second claim, fix $(s, a) \in \mathcal{S} \times \mathcal{A}$. Then

$$\begin{aligned}
&\sum_{s',a'} \big|P_x\big((s,a),(s',a')\big) - P_y\big((s,a),(s',a')\big)\big| \\
&= \sum_{s',a'} p(s' \mid s,a)\,|\chi_x(a' \mid s') - \chi_y(a' \mid s')| \\
&\le \max_{u\in\mathcal{S}} \sum_{b\in\mathcal{A}} |\chi_x(b \mid u) - \chi_y(b \mid u)|.
\end{aligned}$$

Taking the supremum over (s, a) yields the claim. □

Corollary H.3 (Slow variation of the frozen kernels)**.** *There exists a constant $L_{P,\mu} < \infty$, depending only on μ and $(\mathcal{S}, \mathcal{A})$, such that for every $t \ge 1$,*

$$\|P_{t+1} - P_t\|_{1\to 1} \le L_{P,\mu} \frac{\eta_t}{A_{t+1}}.$$

In particular,

$$\|P_{t+1} - P_t\|_{1\to 1} \to 0 \qquad \text{as } t \to \infty.$$

Proof. Apply Lemma H.2 with $x = \bar{x}_{t+1}$ and $y = \bar{x}_t$, and then Lemma H.1. □

H.2. Navigation tracking for the non-homogeneous trajectory

We write

$$\hat{w}_{sa}(t) := \frac{N_{sa}(t)}{t}, \qquad \hat{w}(t) := (\hat{w}_{sa}(t))_{(s,a)\in\mathcal{S}\times\mathcal{A}},$$

for the empirical state-action allocation. For the frozen navigation policies (χ_t), define the tail Cesàro stationary target

$$\vartheta_T := \lceil T^{3/4} \rceil, \qquad \widetilde{\alpha}_T^{\vartheta} := \frac{1}{T - \vartheta_T + 1} \sum_{t=\vartheta_T}^{T} \alpha_p^{\chi_t},$$

and the navigation good event

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

The next proposition is for the non-homogeneous state-action chain generated by the frozen navigation policies. Its proof is inspired by the non-homogeneous Markov-chain framework of (Fort et al., 2011). It combines the common Doeblin minorization from Appendix F, the slow-variation estimate for the frozen kernels (P_t), and a Poisson-equation martingale decomposition tailored to the PT-ACE(μ) trajectory.

Lemma H.4 (Finite-state Poisson equation under a common Doeblin family)**.** *Let Z be a finite state space, and let $\mathfrak{P}$ be a family of stochastic kernels on Z. Assume that there exist an integer $m \ge 1$, a constant $\varepsilon \in (0, 1)$, and a probability distribution ν on Z such that for every $P \in \mathfrak{P}$ and every $z \in \mathsf{Z}$,*

$$P^m(z, \cdot) \ge \varepsilon\, \nu(\cdot).$$

Then the following hold.

(i) Each $P \in \mathfrak{P}$ has a unique invariant distribution α_P. Moreover, there exist constants

$$\rho := (1 - \varepsilon)^{1/m} \in (0, 1), \qquad C_{\mathrm{geo}} := 2\rho^{-(m-1)},$$

such that for every $P \in \mathfrak{P}$, every $k \ge 0$, and every $z \in \mathsf{Z}$,

$$\|P^k(z, \cdot) - \alpha_P\|_1 \le C_{\mathrm{geo}} \rho^k.$$

(ii) For every bounded $f : \mathsf{Z} \to \mathbb{R}$ and every $P \in \mathfrak{P}$, the series

$$g_{P,f} := \sum_{k=0}^{\infty} (P^k - W_P) f, \qquad W_P h := \alpha_P(h)\mathbf{1},$$

converges absolutely in $\ell_\infty(\mathsf{Z})$, satisfies the Poisson equation

$$g_{P,f} - P g_{P,f} = f - \alpha_P(f)\mathbf{1},$$

and obeys the uniform bound

$$\|g_{P,f}\|_\infty \le C_{\mathrm{Pois}} \|f\|_\infty, \qquad C_{\mathrm{Pois}} := \frac{C_{\mathrm{geo}}}{1-\rho}.$$

(iii) There exists a constant $L_{\mathrm{Pois}} < \infty$, depending only on $(m, \varepsilon, |\mathsf{Z}|)$, such that for every bounded $f : \mathsf{Z} \to \mathbb{R}$ and every $P, Q \in \mathfrak{P}$,

$$\|g_{P,f} - g_{Q,f}\|_\infty \le L_{\mathrm{Pois}} \, \|P - Q\|_{1\to 1} \, \|f\|_\infty.$$

Proof. We write W_P for the rank-one kernel whose every row equals α_P, so that

$$(W_P h)(z) = \alpha_P(h), \qquad z \in \mathsf{Z}.$$

Step 1: uniform geometric ergodicity. Fix $P \in \mathfrak{P}$. Since

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

the Dobrushin contraction coefficient of P^m is at most $1 - \varepsilon$. Hence for every signed measure λ on Z with total mass 0,

$$\|\lambda P^m\|_1 \le (1-\varepsilon)\|\lambda\|_1.$$

Iterating gives, for every integer $n \ge 0$,

$$\|\lambda P^{nm}\|_1 \le (1-\varepsilon)^n \|\lambda\|_1.$$

Let α_P be any invariant distribution of P, whose existence is automatic on the finite space Z. Taking $\lambda = \delta_z - \alpha_P$, which has total mass 0, gives

$$\|P^{nm}(z, \cdot) - \alpha_P\|_1 = \|(\delta_z - \alpha_P)P^{nm}\|_1 \le 2(1-\varepsilon)^n.$$

For a general $k \ge 0$, write $k = nm + r$ with $0 \le r \le m - 1$. Since stochastic kernels are contractions in total variation,

$$\|P^k(z, \cdot) - \alpha_P\|_1 = \|(\delta_z - \alpha_P)P^{nm+r}\|_1 \le 2(1-\varepsilon)^n.$$

Using $\rho = (1-\varepsilon)^{1/m}$, we have

$$(1-\varepsilon)^n = \rho^{nm} \le \rho^{k-(m-1)} = \rho^{-(m-1)}\rho^k.$$

Hence

$$\|P^k(z, \cdot) - \alpha_P\|_1 \le 2\rho^{-(m-1)}\rho^k = C_{\mathrm{geo}}\rho^k.$$

In particular, $P^k(z, \cdot)$ converges to the same limit for every z, so α_P is unique.

Step 2: existence and uniform bound for the Poisson solution. Fix bounded f and $P \in \mathfrak{P}$. For every $z \in \mathsf{Z}$,

$$|(P^k - W_P)f(z)| = \big|(P^k(z, \cdot) - \alpha_P)(f)\big| \le \|P^k(z, \cdot) - \alpha_P\|_1 \, \|f\|_\infty \le C_{\mathrm{geo}}\rho^k \|f\|_\infty.$$

Therefore

$$\sum_{k=0}^{\infty} \|(P^k - W_P)f\|_\infty \le \frac{C_{\mathrm{geo}}}{1-\rho}\|f\|_\infty.$$

Thus the series defining $g_{P,f}$ converges absolutely in $\ell_\infty(\mathsf{Z})$, with

$$\|g_{P,f}\|_\infty \le C_{\mathrm{Pois}}\|f\|_\infty.$$

For $N \geq 0$,

$$(I-P)\sum_{k=0}^{N}(P^k - W_P)f = \sum_{k=0}^{N}(P^k - P^{k+1})f = (I - P^{N+1})f.$$

Since $P^{N+1}(z,\cdot) \to \alpha_P$ for every z, we have $P^{N+1}f \to W_P f = \alpha_P(f)\mathbf{1}$ uniformly on the finite space Z. Passing to the limit $N \to \infty$ gives

$$(I-P)g_{P,f} = f - \alpha_P(f)\mathbf{1},$$

which is the desired Poisson equation.

Step 3: Lipschitz continuity of $P \mapsto \alpha_P$. Fix $P, Q \in \mathfrak{P}$, and let

$$\Delta := \alpha_P - \alpha_Q.$$

Since $\Delta(\mathbf{1}) = 0$, we may write

$$\Delta = \alpha_P(P-Q) + \Delta Q.$$

Also, because $\Delta(\mathbf{1}) = 0$,

$$\Delta W_Q = 0, \qquad \Delta Q = \Delta(Q - W_Q).$$

Hence

$$\Delta = \alpha_P(P-Q) + \Delta(Q - W_Q).$$

Iterating this identity yields

$$\Delta = \alpha_P(P-Q)\sum_{k=0}^{\infty}(Q-W_Q)^k.$$

For every signed measure λ with total mass 0,

$$\|\lambda(Q-W_Q)^k\|_1 = \|\lambda Q^k\|_1 \leq C_{\mathrm{geo}}\rho^k\|\lambda\|_1,$$

using Step 1. Therefore

$$\|\alpha_P - \alpha_Q\|_1 \leq \|\alpha_P(P-Q)\|_1 \sum_{k=0}^{\infty} C_{\mathrm{geo}}\rho^k \leq C_{\mathrm{inv}}\|P-Q\|_{1\to 1},$$

where

$$C_{\mathrm{inv}} := \frac{C_{\mathrm{geo}}}{1-\rho}.$$

Since $W_P - W_Q$ has every row equal to $\alpha_P - \alpha_Q$, it follows that

$$\|W_P - W_Q\|_{1\to 1} = \|\alpha_P - \alpha_Q\|_1 \leq C_{\mathrm{inv}}\|P-Q\|_{1\to 1}.$$

Step 4: Lipschitz continuity of the Poisson solution. Define the centered kernels

$$T_P := P - W_P, \qquad T_Q := Q - W_Q,$$

and the centered functions

$$\bar{f}_P := (I - W_P)f = f - \alpha_P(f)\mathbf{1}, \qquad \bar{f}_Q := (I - W_Q)f = f - \alpha_Q(f)\mathbf{1}.$$

Since $PW_P = W_P$, $W_P P = W_P$, and $W_P^2 = W_P$, we have

$$T_P^k = P^k - W_P, \qquad k \geq 1,$$

and therefore

$$g_{P,f} = \sum_{k=0}^{\infty} T_P^k \bar{f}_P.$$

Similarly,

$$g_{Q,f} = \sum_{k=0}^{\infty} T_Q^k \bar{f}_Q.$$

Hence

$$g_{P,f} - g_{Q,f} = \sum_{k=0}^{\infty} (T_P^k - T_Q^k)\bar{f}_P + \sum_{k=0}^{\infty} T_Q^k(\bar{f}_P - \bar{f}_Q).$$

We bound the two terms separately.

First, for $k \geq 1$,

$$T_P^k - T_Q^k = \sum_{j=0}^{k-1} T_P^j (T_P - T_Q) T_Q^{k-1-j}.$$

By Step 1,

$$\|T_P^j h\|_\infty = \|(P^j - W_P)h\|_\infty \leq C_{\text{geo}} \rho^j \|h\|_\infty, \qquad j \geq 0,$$

and similarly for T_Q. Moreover,

$$\|T_P - T_Q\|_{1\to 1} \leq \|P - Q\|_{1\to 1} + \|W_P - W_Q\|_{1\to 1} \leq (1 + C_{\text{inv}})\|P - Q\|_{1\to 1}.$$

Since $\|\bar{f}_P\|_\infty \leq 2\|f\|_\infty$, we obtain for $k \geq 1$,

$$\begin{aligned}\|(T_P^k - T_Q^k)\bar{f}_P\|_\infty &\leq \sum_{j=0}^{k-1} \|T_P^j\|_{\infty\to\infty} \|T_P - T_Q\|_{1\to 1} \|T_Q^{k-1-j} \bar{f}_P\|_\infty \\ &\leq 2k\, C_{\text{geo}}^2 \rho^{k-1} (1 + C_{\text{inv}}) \|P - Q\|_{1\to 1} \|f\|_\infty.\end{aligned}$$

Summing over $k \geq 1$ gives

$$\left\| \sum_{k=0}^{\infty} (T_P^k - T_Q^k)\bar{f}_P \right\|_\infty \leq \frac{2C_{\text{geo}}^2 (1 + C_{\text{inv}})}{(1-\rho)^2} \|P - Q\|_{1\to 1} \|f\|_\infty.$$

For the second term, note that

$$\bar{f}_P - \bar{f}_Q = (W_Q - W_P) f.$$

Hence

$$\|\bar{f}_P - \bar{f}_Q\|_\infty \leq \|W_P - W_Q\|_{1\to 1} \|f\|_\infty \leq C_{\text{inv}} \|P - Q\|_{1\to 1} \|f\|_\infty.$$

Using again the uniform bound on T_Q^k,

$$\left\| \sum_{k=0}^{\infty} T_Q^k (\bar{f}_P - \bar{f}_Q) \right\|_\infty \leq \frac{C_{\text{geo}} C_{\text{inv}}}{1-\rho} \|P - Q\|_{1\to 1} \|f\|_\infty.$$

Combining the two bounds yields the claim, with

$$L_{\text{Pois}} := \frac{2C_{\text{geo}}^2 (1 + C_{\text{inv}})}{(1-\rho)^2} + \frac{C_{\text{geo}} C_{\text{inv}}}{1-\rho}.$$

□

Proposition H.5 (Quantitative navigation probability bound)**.** *Assume Assumption 3.6. Fix $\zeta > 0$. Then there exist constants*

$$B_{\text{nav}}(\mu, p, \zeta) > 0, \qquad C_{\text{nav}}(\mu, p, \zeta) > 0,$$

such that for all sufficiently large T,

$$\mathbb{P}((\mathcal{C}_T^{\text{nav}}(\zeta))^c) \leq B_{\text{nav}}(\mu, p, \zeta) \exp\bigl(-C_{\text{nav}}(\mu, p, \zeta) T\bigr).$$

Proof. Fix $\zeta > 0$.

Step 1: uniform ergodicity and slow variation of the frozen kernels. Let $Z_t := (S_t, A_t)$, and let P_t denote the predictable state-action transition kernel governing $Z_t \to Z_{t+1}$. By construction of the frozen policies and the fact that $\bar{x}_t \in K_\mu$, we have

$$\chi_t(a \mid s) = \frac{\bar{x}_t(s,a)}{\sum_{b\in\mathcal{A}} \bar{x}_t(s,b)} \geq \bar{x}_t(s,a) \geq \mu, \qquad \forall (s,a) \in \mathcal{S}\times\mathcal{A},\ \forall t \geq 1.$$

Hence, if

$$P_u := P_p^{\mathrm{Unif}_{\mathcal{A}}}$$

denotes the state-action kernel induced by the uniform policy, then

$$P_t \geq |\mathcal{A}|\mu\, P_u \qquad \text{entrywise for all } t \geq 1.$$

By Assumption 3.6, P_u is irreducible and aperiodic on the finite space $\mathcal{S}\times\mathcal{A}$. Therefore there exist an integer $m_\mu \geq 1$, a constant $\varepsilon_\mu \in (0,1)$, and a probability distribution ν_μ on $\mathcal{S}\times\mathcal{A}$, depending only on (μ, p), such that every frozen kernel P_t satisfies

$$P_t^{m_\mu}(z,\cdot) \geq \varepsilon_\mu \nu_\mu(\cdot), \qquad \forall z \in \mathcal{S}\times\mathcal{A}.$$

This gives a common Doeblin family for the frozen kernels.

On the other hand, by Corollary H.3,

$$\|P_{t+1} - P_t\|_{1\to 1} \leq L_{P,\mu}\frac{\eta_t}{A_{t+1}}, \qquad \forall t \geq 1.$$

Since $A_t \to \infty$ and $\eta_t/A_{t+1} \to 0$, the sequence (P_t) is asymptotically stable.

Step 2: Poisson equation and martingale decomposition. Define the uniformly minorized family

$$\mathfrak{P}_\mu := \Big\{P^\chi :\ \chi(a \mid s) \geq \mu\ \forall (s,a) \in \mathcal{S}\times\mathcal{A}\Big\}.$$

Each frozen kernel P_t belongs to $\mathfrak{P}_\mu$. Hence Lemma H.4 applies uniformly on $\mathfrak{P}_\mu$.

Fix any bounded function $f : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$. For each $t \geq 1$, let $\alpha_t := \alpha_p^{\chi_t}$ denote the invariant distribution of P_t, and let

$$g_t := g_{P_t,f}$$

be the Poisson solution associated with P_t and f. Then

$$g_t - P_t g_t = f - \alpha_t(f),$$

and there exist constants $C_{g,\mu,p}, C'_{g,\mu,p} < \infty$, depending only on (μ, p), such that

$$\|g_t\|_\infty \leq C_{g,\mu,p}\|f\|_\infty,$$

and

$$\|g_{t+1} - g_t\|_\infty \leq C'_{g,\mu,p}\|P_{t+1} - P_t\|_{1\to 1}\|f\|_\infty, \qquad \forall t \geq 1.$$

Define

$$M_n(f) := \sum_{t=1}^{n-1}\Big(g_t(Z_{t+1}) - P_t g_t(Z_t)\Big).$$

Since P_t is the predictable kernel for $Z_t \to Z_{t+1}$, $(M_n(f))_{n\geq 1}$ is a martingale with respect to the natural filtration of (Z_t). Its increments satisfy

$$|M_{t+1}(f) - M_t(f)| = \big|g_t(Z_{t+1}) - P_t g_t(Z_t)\big| \leq 2C_{g,\mu,p}\|f\|_\infty.$$

Using $f - \alpha_t(f) = g_t - P_t g_t$, a telescoping-sum computation gives, for every $1 \leq a \leq T$,

$$\sum_{t=a}^{T}\big(f(Z_t) - \alpha_t(f)\big) = M_{T+1}(f) - M_a(f) + g_a(Z_a) - g_T(Z_{T+1}) + \sum_{t=a+1}^{T}\big(g_t - g_{t-1}\big)(Z_t). \tag{9}$$

Taking absolute values and using the bounds above yields

$$\left|\sum_{t=a}^{T}\big(f(Z_t) - \alpha_t(f)\big) - \big(M_{T+1}(f) - M_a(f)\big)\right| \leq 2C_{g,\mu,p}\|f\|_\infty + C'_{g,\mu,p}\|f\|_\infty \sum_{t=a+1}^{T}\|P_t - P_{t-1}\|_{1\to 1}. \tag{10}$$

Step 3: specialization to coordinate indicators. Fix $(s,a) \in \mathcal{S} \times \mathcal{A}$, and let

$$f_{sa}(z) := \mathbf{1}\{z = (s,a)\}.$$

For this choice,

$$\frac{1}{T}\sum_{t=1}^{T} f_{sa}(Z_t) = \hat{w}_{sa}(T).$$

Apply (9)–(10) with $a = \vartheta_T$, where $\vartheta_T = \lceil T^{3/4} \rceil$. Since

$$\widetilde{\alpha}_T^{\vartheta}(s,a) = \frac{1}{T - \vartheta_T + 1}\sum_{t=\vartheta_T}^{T} \alpha_t(s,a),$$

we obtain

$$\left| \frac{1}{T - \vartheta_T + 1}\sum_{t=\vartheta_T}^{T} f_{sa}(Z_t) - \widetilde{\alpha}_T^{\vartheta}(s,a) \right| \le \frac{|M_{T+1}(f_{sa}) - M_{\vartheta_T}(f_{sa})|}{T - \vartheta_T + 1} + \frac{2C_{g,\mu,p} + C'_{g,\mu,p}\sum_{t=\vartheta_T+1}^{T} \|P_t - P_{t-1}\|_{1\to 1}}{T - \vartheta_T + 1}.$$

By Corollary H.3,

$$\sum_{t=\vartheta_T+1}^{T} \|P_t - P_{t-1}\|_{1\to 1} \le L_{P,\mu} \sum_{t=\vartheta_T+1}^{T} \frac{\eta_{t-1}}{A_t}.$$

Since $A_t = A_{t-1} + \eta_{t-1}$,

$$\sum_{t=\vartheta_T+1}^{T} \frac{\eta_{t-1}}{A_t} \le \log A_T - \log A_{\vartheta_T}.$$

Therefore

$$\frac{2C_{g,\mu,p} + C'_{g,\mu,p}\sum_{t=\vartheta_T+1}^{T} \|P_t - P_{t-1}\|_{1\to 1}}{T - \vartheta_T + 1} = o(1).$$

Hence, for all sufficiently large T, the deterministic remainder above is at most

$$\frac{\zeta}{4|\mathcal{S}||\mathcal{A}|}.$$

Step 4: concentration of the martingale term. Since the martingale increments are bounded by $2C_{g,\mu,p}$, Azuma–Hoeffding yields

$$\mathbb{P}\left(|M_{T+1}(f_{sa}) - M_{\vartheta_T}(f_{sa})| > \frac{\zeta}{4|\mathcal{S}||\mathcal{A}|}(T - \vartheta_T + 1) \right) \le 2\exp\left(-\frac{\zeta^2 (T - \vartheta_T + 1)}{32|\mathcal{S}|^2|\mathcal{A}|^2 C_{g,\mu,p}^2} \right).$$

Hence there exists $c_{\mu,p,\zeta} > 0$ such that for all sufficiently large T,

$$\mathbb{P}\left(\frac{|M_{T+1}(f_{sa}) - M_{\vartheta_T}(f_{sa})|}{T - \vartheta_T + 1} > \frac{\zeta}{4|\mathcal{S}||\mathcal{A}|} \right) \le 2\exp\bigl(-c_{\mu,p,\zeta} T\bigr).$$

Therefore, for every $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\mathbb{P}\left(\left| \frac{1}{T - \vartheta_T + 1}\sum_{t=\vartheta_T}^{T} f_{sa}(Z_t) - \widetilde{\alpha}_T^{\vartheta}(s,a) \right| > \frac{\zeta}{2|\mathcal{S}||\mathcal{A}|} \right) \le 2\exp\bigl(-c_{\mu,p,\zeta} T\bigr).$$

Step 5: return to $\hat{w}(T)$. It remains to replace the tail empirical average by $\hat{w}(T)$. Since $0 \le f_{sa} \le 1$,

$$\left|\hat{w}_{sa}(T) - \frac{1}{T-\vartheta_T+1}\sum_{t=\vartheta_T}^{T} f_{sa}(Z_t)\right| \le \frac{\vartheta_T - 1}{T} + \left|\frac{1}{T} - \frac{1}{T-\vartheta_T+1}\right| \sum_{t=\vartheta_T}^{T} f_{sa}(Z_t),$$

and hence

$$\left|\hat{w}_{sa}(T) - \frac{1}{T-\vartheta_T+1}\sum_{t=\vartheta_T}^{T} f_{sa}(Z_t)\right| \le 2\frac{\vartheta_T}{T}.$$

Because $\vartheta_T = \lceil T^{3/4} \rceil$, we have $2\vartheta_T/T = o(1)$. Thus, for all sufficiently large T,

$$\left|\hat{w}_{sa}(T) - \widetilde{\alpha}_T^{\vartheta}(s,a)\right| \le \left|\frac{1}{T-\vartheta_T+1}\sum_{t=\vartheta_T}^{T} f_{sa}(Z_t) - \widetilde{\alpha}_T^{\vartheta}(s,a)\right| + \frac{\zeta}{2|\mathcal{S}||\mathcal{A}|}.$$

Therefore

$$\mathbb{P}\left(\left|\hat{w}_{sa}(T) - \widetilde{\alpha}_T^{\vartheta}(s,a)\right| > \frac{\zeta}{|\mathcal{S}||\mathcal{A}|}\right) \le 2\exp\bigl(-c_{\mu,p,\zeta}T\bigr).$$

Taking a union bound over $(s,a) \in \mathcal{S} \times \mathcal{A}$, we obtain

$$\mathbb{P}\left(\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|\hat{w}_{sa}(T) - \widetilde{\alpha}_T^{\vartheta}(s,a)\right| > \zeta\right) \le 2|\mathcal{S}||\mathcal{A}|\,\exp\bigl(-c_{\mu,p,\zeta}T\bigr).$$

Since the event inside the probability is exactly $(\mathcal{C}_T^{\mathrm{nav}}(\zeta))^c$, the claimed bound follows with

$$B_{\mathrm{nav}}(\mu,p,\zeta) := 2|\mathcal{S}||\mathcal{A}|, \qquad C_{\mathrm{nav}}(\mu,p,\zeta) := c_{\mu,p,\zeta}.$$

□

H.3. Approximate flow implies occupancy tracking

The next lemma is the deterministic bridge between the exchange allocations $\bar{x}_t$ analyzed in Appendix G and the stationary occupancies $\alpha_p^{\chi_t}$ of the frozen navigation policies.

Lemma H.6 (Approximate flow implies occupancy tracking)**.** *Fix $\mu > 0$, and define*

$$K_\mu := \{x \in \Delta_{\mathcal{S}\times\mathcal{A}} : x_{sa} \ge \mu\ \forall (s,a)\}.$$

For $x \in K_\mu$, define

$$x_s := \sum_{a\in\mathcal{A}} x_{sa}, \qquad \chi_x(a \mid s) := \frac{x_{sa}}{x_s}.$$

Assume Assumption 3.6. Then there exists a constant $C_\mu(p) < \infty$, depending only on (μ, p), such that for every $x \in K_\mu$,

$$\|\alpha_p^{\chi_x} - x\|_1 \le C_\mu(p) \sum_{s\in\mathcal{S}} \left| x_s - \sum_{y\in\mathcal{S}, b\in\mathcal{A}} x_{yb}\, p(s \mid y, b) \right|.$$

Proof. For $x \in K_\mu$, let P_x be the state-action kernel induced by (p, χ_x):

$$P_x\bigl((y,b),(s,a)\bigr) := p(s \mid y,b)\chi_x(a \mid s).$$

Then $\alpha_p^{\chi_x}$ is the invariant distribution of P_x.

For every $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$(xP_x)_{sa} = \sum_{y\in\mathcal{S}, b\in\mathcal{A}} x_{yb}\, p(s \mid y,b)\chi_x(a \mid s) = \chi_x(a \mid s) \sum_{y\in\mathcal{S}, b\in\mathcal{A}} x_{yb}\, p(s \mid y,b),$$

whereas

$$x_{sa} = x_s \chi_x(a \mid s).$$

Hence

$$x_{sa} - (xP_x)_{sa} = \chi_x(a \mid s) \left(x_s - \sum_{y \in \mathcal{S}, b \in \mathcal{A}} x_{yb} p(s \mid y, b) \right).$$

Summing over $a \in \mathcal{A}$ gives

$$\sum_{a \in \mathcal{A}} |x_{sa} - (xP_x)_{sa}| = \left| x_s - \sum_{y \in \mathcal{S}, b \in \mathcal{A}} x_{yb} p(s \mid y, b) \right|.$$

Thus

$$\|x - xP_x\|_1 = \sum_{s \in \mathcal{S}} \left| x_s - \sum_{y \in \mathcal{S}, b \in \mathcal{A}} x_{yb} p(s \mid y, b) \right|. \tag{11}$$

We next show that the family $(P_x)_{x \in K_\mu}$ is uniformly ergodic. Since $x_{sa} \geq \mu$ for all (s, a), we have $x_s \leq 1$, and therefore

$$\chi_x(a \mid s) = \frac{x_{sa}}{x_s} \geq \mu.$$

Hence, entrywise,

$$P_x\big((y, b), (s, a)\big) = p(s \mid y, b)\chi_x(a \mid s) \geq |\mathcal{A}|\mu \, P_p^{\mathrm{Unif}_{\mathcal{A}}}\big((y, b), (s, a)\big).$$

By Assumption 3.6, the state-action chain under $P_p^{\mathrm{Unif}_{\mathcal{A}}}$ is irreducible and aperiodic. Since $\mathcal{S} \times \mathcal{A}$ is finite, there exist $m \geq 1$ and $\varepsilon_\mu > 0$ such that

$$P_x^m(z, \cdot) \geq \varepsilon_\mu u(\cdot) \qquad \forall z \in \mathcal{S} \times \mathcal{A}, \ \forall x \in K_\mu,$$

where u is the uniform distribution on $\mathcal{S} \times \mathcal{A}$. Thus, for every signed measure ν with total mass 0,

$$\|\nu P_x^{nm}\|_1 \leq (1 - \varepsilon_\mu)^n \|\nu\|_1, \qquad n \geq 0. \tag{12}$$

Since x and $\alpha_p^{\chi_x}$ are probability measures and $\alpha_p^{\chi_x}$ is invariant for P_x,

$$x - \alpha_p^{\chi_x} = \lim_{M \to \infty} (x - xP_x) \sum_{k=0}^{M} P_x^k.$$

Taking ℓ_1-norms, decomposing $k = nm + j$, and using (12), we obtain

$$\|x - \alpha_p^{\chi_x}\|_1 \leq \frac{m}{\varepsilon_\mu} \|x - xP_x\|_1.$$

Combining this with (11) proves the lemma with

$$C_\mu(p) := \frac{m}{\varepsilon_\mu}.$$

□

Corollary H.7 (Tail-uniform occupancy tracking from approximate flow)**.** *Suppose there exists a deterministic quantity* $r_{\mathrm{flow}}(T, \zeta) \geq 0$ *such that on* $\mathcal{C}_T^{\mathrm{ker}}(\zeta)$*,*

$$\sup_{t \in [\tau_T, T]} \sum_{s \in \mathcal{S}} \left| \bar{x}_{t,s} - \sum_{y \in \mathcal{S}, b \in \mathcal{A}} \bar{x}_t(y, b) p(s \mid y, b) \right| \leq r_{\mathrm{flow}}(T, \zeta).$$

Then, on $\mathcal{C}_T^{\mathrm{ker}}(\zeta)$*,*

$$\sup_{t \in [\tau_T, T]} \|\alpha_p^{\chi_t} - \bar{x}_t\|_1 \leq C_\mu(p) \, r_{\mathrm{flow}}(T, \zeta).$$

Proof. For every $t \in [\tau_T, T]$, we have $\bar{x}_t \in K_\mu$, since each exchange iterate lies in K_μ and K_μ is convex. Also, by definition of the frozen navigation rule,

$$\chi_t = \chi_{\bar{x}_t}.$$

Applying Lemma H.6 with $x = \bar{x}_t$ gives

$$\|\alpha_p^{\chi_t} - \bar{x}_t\|_1 \le C_\mu(p) \sum_{s \in \mathcal{S}} \left| \bar{x}_{t,s} - \sum_{y \in \mathcal{S}, b \in \mathcal{A}} \bar{x}_t(y,b) p(s \mid y,b) \right|.$$

Taking the supremum over $t \in [\tau_T, T]$ yields the claim. □

H.4. Consequence of the enhanced exchange analysis

We now combine the navigation event with the tail-window conclusions already established in Appendix G, namely Proposition G.16 and Proposition G.17.

Define

$$\vartheta_T := \lceil T^{3/4} \rceil, \qquad \widetilde{\alpha}_T^{\vartheta} := \frac{1}{T - \vartheta_T + 1} \sum_{t=\vartheta_T}^{T} \alpha_p^{\chi_t}, \qquad \widetilde{x}_T^{\vartheta} := \frac{1}{T - \vartheta_T + 1} \sum_{t=\vartheta_T}^{T} \bar{x}_t.$$

Lemma H.8 (Uniform interiority of navigation occupancies)**.** *Assume Assumption 3.6, and let* $0 < \mu \le 1/|\mathcal{A}|$*. Then there exists a constant*

$$\underline{\alpha}_{\text{nav}}(p, \mu) > 0$$

such that for every stationary policy χ *satisfying*

$$\chi(a \mid s) \ge \mu \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

the stationary state-action occupancy α_p^χ *satisfies*

$$\alpha_p^\chi(s,a) \ge \underline{\alpha}_{\text{nav}}(p, \mu) \qquad \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

In particular,

$$\alpha_p^\chi \in K_{\underline{\alpha}_{\text{nav}}(p,\mu)}.$$

Consequently, any convex average of such occupancies also belongs to $K_{\underline{\alpha}_{\text{nav}}(p,\mu)}$.

Proof. Let $P^u := P_p^{\text{Unif}_{\mathcal{A}}}$ denote the state-action transition kernel induced by the uniform policy:

$$P^u\big((s,a),(s',a')\big) = p(s' \mid s,a)\text{Unif}_{\mathcal{A}}(a' \mid s').$$

By Assumption 3.6, P^u is irreducible and aperiodic on the finite space $\mathcal{S} \times \mathcal{A}$. Hence there exists an integer $m \ge 1$ such that

$$(P^u)^m(z,z') > 0 \qquad \forall z, z' \in \mathcal{S} \times \mathcal{A}.$$

Define

$$\delta_u := \min_{z,z'} (P^u)^m(z,z') > 0.$$

Now let χ be any stationary policy satisfying $\chi(a \mid s) \ge \mu$ for all (s,a). Let $P^\chi := P_p^\chi$ be the induced state-action transition kernel:

$$P^\chi\big((s,a),(s',a')\big) = p(s' \mid s,a)\chi(a' \mid s').$$

Then for every $z = (s,a)$ and $z' = (s',a')$,

$$P^\chi(z,z') = p(s' \mid s,a)\chi(a' \mid s') \ge \mu\, p(s' \mid s,a) = \mu|\mathcal{A}|\, P^u(z,z').$$

Since matrix multiplication preserves entrywise inequalities for nonnegative matrices,

$$(P^\chi)^m \ge (\mu|\mathcal{A}|)^m (P^u)^m$$

entrywise. In particular,

$$(P^\chi)^m(z,z') \geq (\mu|\mathcal{A}|)^m \delta_u \qquad \forall z, z'.$$

Because $(P^\chi)^m$ is strictly positive, P^χ is irreducible and aperiodic, and thus admits a unique stationary distribution α_p^χ. Using stationarity,

$$\alpha_p^\chi = \alpha_p^\chi (P^\chi)^m.$$

Therefore, for every $z' \in \mathcal{S} \times \mathcal{A}$,

$$\alpha_p^\chi(z') = \sum_z \alpha_p^\chi(z)\,(P^\chi)^m(z,z') \geq (\mu|\mathcal{A}|)^m \delta_u \sum_z \alpha_p^\chi(z) = (\mu|\mathcal{A}|)^m \delta_u.$$

Hence the claim holds with

$$\underline{\alpha}_{\mathrm{nav}}(p,\mu) := (\mu|\mathcal{A}|)^m \delta_u > 0.$$

The final claim follows because $K_{\underline{\alpha}_{\mathrm{nav}}(p,\mu)}$ is convex. □

Proposition H.9 (Navigation reduction)**.** *Assume Assumption 2.2 holds and $p \in \mathcal{P}_{\mathrm{test}}^{1:J}$. Fix an admissible $\mu > 0$ and $\zeta > 0$. Let*

$$\underline{\alpha}_{\mathrm{nav}}(p,\mu) > 0$$

be the constant from Lemma H.8, and define

$$\mu_\star := \frac{1}{2}\min\{\mu, \underline{\alpha}_{\mathrm{nav}}(p,\mu)\}.$$

Let $\psi_{p,\mu_\star}$ be a continuity modulus of $F_p^{1:J}$ on the compact set $K_{\mu_\star}$.

Then, for every $0 < \zeta \leq \underline{\alpha}_{\mathrm{nav}}(p,\mu)/2$, on

$$\mathcal{C}_T^{\mathrm{ker}}(\zeta) \cap \mathcal{C}_T^{\mathrm{nav}}(\zeta),$$

for all sufficiently large T,

$$F_p^{1:J}(\hat{w}(T)) \geq \frac{1}{1+c(\mu,p)} \max_{w \in \Sigma_p} F_p^{1:J}(w) - r_{\mathrm{ex}}^{\mathrm{tail}}(T,\mu,\zeta) - \psi_{p,\mu_\star}\Big(\zeta + C_\mu(p)\, r_{\mathrm{flow}}(T,\zeta)\Big).$$

Consequently,

$$\lim_{\zeta \downarrow 0} \liminf_{T\to\infty} F_p^{1:J}(\hat{w}(T)) \geq \frac{1}{1+c(\mu,p)} \max_{w \in \Sigma_p} F_p^{1:J}(w).$$

Proof. Work on $\mathcal{C}_T^{\mathrm{ker}}(\zeta) \cap \mathcal{C}_T^{\mathrm{nav}}(\zeta)$.

By Proposition G.17 and Lemma H.6, for every $t \in [\vartheta_T, T]$,

$$\|\alpha_p^{\chi_t} - \bar{x}_t\|_1 \leq C_\mu(p) \sum_{s\in\mathcal{S}} \left| \bar{x}_{t,s} - \sum_{y\in\mathcal{S}, b\in\mathcal{A}} \bar{x}_t(y,b) p(s \mid y,b) \right| \leq C_\mu(p)\, r_{\mathrm{flow}}(T,\zeta).$$

Averaging over $t = \vartheta_T, \ldots, T$, we obtain

$$\|\widetilde{\alpha}_T^\vartheta - \widetilde{x}_T^\vartheta\|_1 \leq C_\mu(p)\, r_{\mathrm{flow}}(T,\zeta). \tag{13}$$

Because $\mathcal{C}_T^{\mathrm{nav}}(\zeta)$ holds,

$$\|\hat{w}(T) - \widetilde{\alpha}_T^\vartheta\|_1 \leq \zeta.$$

Combining this with (13), we get

$$\|\hat{w}(T) - \widetilde{x}_T^\vartheta\|_1 \leq \zeta + C_\mu(p)\, r_{\mathrm{flow}}(T,\zeta). \tag{14}$$

We next verify that the relevant points lie in a common compact interior slice. Since each $\bar{x}_t \in K_\mu$ and K_μ is convex, we have

$$\widetilde{x}_T^\vartheta \in K_\mu.$$

Also, by Lemma H.8, each $\alpha_p^{\chi_t}$ belongs to $K_{\underline{\alpha}_{\mathrm{nav}}(p,\mu)}$. Since this set is convex,

$$\widetilde{\alpha}_T^{\vartheta} \in K_{\underline{\alpha}_{\mathrm{nav}}(p,\mu)}.$$

Because $\mathcal{C}_T^{\mathrm{nav}}(\zeta)$ holds and $0 < \zeta \le \underline{\alpha}_{\mathrm{nav}}(p,\mu)/2$, for every $(s,a) \in \mathcal{S} \times \mathcal{A}$,

$$\hat{w}_{sa}(T) \ge \widetilde{\alpha}_{T,sa}^{\vartheta} - \zeta \ge \underline{\alpha}_{\mathrm{nav}}(p,\mu) - \zeta \ge \frac{1}{2}\,\underline{\alpha}_{\mathrm{nav}}(p,\mu).$$

Hence

$$\hat{w}(T) \in K_{\underline{\alpha}_{\mathrm{nav}}(p,\mu)/2}.$$

Therefore both $\hat{w}(T)$ and $\widetilde{x}_T^{\vartheta}$ belong to

$$K_{\mu_\star}, \qquad \mu_\star := \frac{1}{2}\min\{\mu, \underline{\alpha}_{\mathrm{nav}}(p,\mu)\}.$$

We now lower bound $F_p^{1:J}(\widetilde{x}_T^{\vartheta})$. By concavity of $F_p^{1:J}$,

$$F_p^{1:J}(\widetilde{x}_T^{\vartheta}) = F_p^{1:J}\left(\frac{1}{T-\vartheta_T+1}\sum_{t=\vartheta_T}^{T}\bar{x}_t\right) \ge \frac{1}{T-\vartheta_T+1}\sum_{t=\vartheta_T}^{T} F_p^{1:J}(\bar{x}_t).$$

By Proposition G.16, on $\mathcal{C}_T^{\mathrm{ker}}(\zeta)$,

$$\inf_{t\in[\vartheta_T,T]} F_p^{1:J}(\bar{x}_t) \ge \frac{1}{1+c(\mu,p)}\max_{w\in\Sigma_p} F_p^{1:J}(w) - r_{\mathrm{ex}}^{\mathrm{tail}}(T,\mu,\zeta).$$

Therefore

$$F_p^{1:J}(\widetilde{x}_T^{\vartheta}) \ge \frac{1}{1+c(\mu,p)}\max_{w\in\Sigma_p} F_p^{1:J}(w) - r_{\mathrm{ex}}^{\mathrm{tail}}(T,\mu,\zeta). \tag{15}$$

Finally, since both $\hat{w}(T)$ and $\widetilde{x}_T^{\vartheta}$ lie in the compact set $K_{\mu_\star}$, the modulus $\psi_{p,\mu_\star}$ yields, together with (14),

$$F_p^{1:J}(\hat{w}(T)) \ge F_p^{1:J}(\widetilde{x}_T^{\vartheta}) - \psi_{p,\mu_\star}\Big(\zeta + C_\mu(p)\, r_{\mathrm{flow}}(T,\zeta)\Big).$$

Combining this with (15) proves the stated lower bound.

For the limiting claim, take the lower bound above, then first take $\liminf_{T\to\infty}$ and then send $\zeta \downarrow 0$. By Proposition G.16,

$$\lim_{\zeta\downarrow 0}\limsup_{T\to\infty} r_{\mathrm{ex}}^{\mathrm{tail}}(T,\mu,\zeta) = 0,$$

and by Proposition G.17,

$$\lim_{\zeta\downarrow 0}\limsup_{T\to\infty} r_{\mathrm{flow}}(T,\zeta) = 0.$$

Since $\psi_{p,\mu_\star}(r) \to 0$ as $r \downarrow 0$, the claimed two-stage limit follows. □

I. Sensitivity of the stopping frontier on a compact interior slice

In this appendix we prove the stopping-side sensitivity statement needed later in the proof of the final upper bound. The key point is that the stopping oracle is queried at the empirical allocation $\hat{w}(T)$, and thus we first show that, on the navigation good event, $\hat{w}(T)$ eventually belongs to a fixed compact interior slice. We then prove a uniform local separation statement for the policywise testing frontiers on that compact slice.

Support-safe continuity on a compact interior slice. All continuity statements in this appendix are taken on a compact interior slice $K_{\mu_{\mathrm{nav}}} \subset \Delta(\mathcal{S} \times \mathcal{A})$, with $\mu_{\mathrm{nav}} > 0$, and for kernels p' in a sufficiently small neighborhood of the true kernel p. This restriction is essential: KL-based objectives need not be continuous on the boundary of the simplex. The arguments below therefore never rely on ambient continuity of $(p', w, q) \mapsto \Phi(p'; w, q)$ on the full product $\mathcal{P} \times \Delta(\mathcal{S} \times \mathcal{A}) \times \mathcal{P}$. Instead, continuity is used only at support-safe points, and the required support-safe approximations are obtained from the policywise closed/strict frontier equivalence recorded in Appendix C.

I.1. Eventual interiority of the empirical allocation

Recall from Appendix H the tail Cesàro stationary target

$$\vartheta_T := \lceil T^{3/4} \rceil, \qquad \widetilde{\alpha}_T^{\vartheta} := \frac{1}{T - \vartheta_T + 1} \sum_{t=\vartheta_T}^{T} \alpha_p^{\chi_t},$$

where χ_t is the frozen navigation policy governing $Z_t \to Z_{t+1}$. The navigation event is

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

and the combined good event is

$$\mathcal{C}_T(\zeta) := \mathcal{C}_T^{\mathrm{ker}}(\zeta) \cap \mathcal{C}_T^{\mathrm{nav}}(\zeta).$$

Lemma I.1 (Uniform interiority of the stationary occupancies)**.** *Assume Assumption 3.6 and fix $\mu > 0$. Then there exists a constant*

$$\underline{\alpha}_{\mu,p} > 0$$

such that for every $t \ge 1$,

$$\alpha_p^{\chi_t}(s, a) \ge \underline{\alpha}_{\mu,p} \qquad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Consequently, for every $T \ge 1$,

$$\widetilde{\alpha}_T^{\vartheta}(s, a) \ge \underline{\alpha}_{\mu,p} \qquad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Proof. This is Lemma H.8 from Appendix H, specialized to the frozen navigation policies (χ_t). The second claim follows by averaging over $t = \vartheta_T, \ldots, T$. □

Corollary I.2 (Eventual interiority of the empirical allocation)**.** *Define*

$$\mu_{\mathrm{nav}} := \frac{\underline{\alpha}_{\mu,p}}{2}.$$

If $0 < \zeta \le \underline{\alpha}_{\mu,p}/2$, then on $\mathcal{C}_T^{\mathrm{nav}}(\zeta)$,

$$\hat{w}(T) \in K_{\mu_{\mathrm{nav}}}.$$

Equivalently,

$$\hat{w}_{sa}(T) \ge \mu_{\mathrm{nav}} \qquad \forall (s, a) \in \mathcal{S} \times \mathcal{A}.$$

Proof. Fix $(s, a) \in \mathcal{S} \times \mathcal{A}$. On $\mathcal{C}_T^{\mathrm{nav}}(\zeta)$,

$$\|\hat{w}(T) - \widetilde{\alpha}_T^{\vartheta}\|_1 \le \zeta,$$

and hence, coordinatewise,

$$\hat{w}_{sa}(T) \ge \widetilde{\alpha}_T^{\vartheta}(s, a) - \zeta.$$

By Lemma I.1,

$$\widetilde{\alpha}_T^{\vartheta}(s, a) \ge \underline{\alpha}_{\mu,p}.$$

Therefore

$$\hat{w}_{sa}(T) \geq \underline{\alpha}_{\mu,p} - \zeta \geq \frac{\underline{\alpha}_{\mu,p}}{2} = \mu_{\text{nav}}.$$

Since this holds for every (s, a), we conclude that

$$\hat{w}(T) \in K_{\mu_{\text{nav}}}.$$

□

I.2. Uniform local policywise testing separation on $K_{\mu_{\text{nav}}}$

Fix a policy index $j \in [J]$. For $p' \in \mathcal{P}$, $w \in K_{\mu_{\text{nav}}}$, and $q \in \mathcal{P}$, define

$$g_{p',j}(q) := V_{p',j}(\rho)V_{q,j}(\rho), \qquad \Phi(p'; w, q) := \sum_{s,a} w_{sa}\text{KL}_{sa}(p', q).$$

The policywise testing frontier value is

$$u_{\text{NO}}^{(j)}(\sigma, w, p') := \min\big\{g_{p',j}(q) : \Phi(p'; w, q) \leq \sigma,\ q \in \mathcal{P}\big\}, \qquad \sigma \geq 0.$$

We also define the strict and closed policywise testing root values

$$\sigma_{\text{NC}}^{(j)}(0, w, p') := \inf\big\{\Phi(p'; w, q) : g_{p',j}(q) < 0,\ q \in \mathcal{P}\big\},$$

and

$$\bar{\sigma}_j(w, p') := \min\big\{\Phi(p'; w, q) : g_{p',j}(q) \leq 0,\ q \in \mathcal{P}\big\}.$$

On the local testing neighborhood below, these two root values agree by the policywise frontier facts in Appendix C.

Lemma I.3 (Local sign stability around p)**.** *Assume $p \in \mathcal{P}_{\text{test}}^{1:J}$. Then there exist constants*

$$a_p > 0, \qquad \zeta_p > 0,$$

such that for every $p' \in \mathcal{P}$ with

$$\|p' - p\|_\infty \leq \zeta_p,$$

and every $j \in [J]$, one has

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

In particular, every such p' belongs to $\mathcal{P}_{\text{test}}^{1:J}$.

Proof. Since $p \in \mathcal{P}_{\text{test}}^{1:J}$, we have $V_{p,j}(\rho) \neq 0$ for every $j \in [J]$. By continuity of $p' \mapsto V_{p',j}(\rho)$, for each j there exists $\zeta_{p,j} > 0$ such that

$$\|p' - p\|_\infty \leq \zeta_{p,j} \quad \Longrightarrow \quad |V_{p',j}(\rho) - V_{p,j}(\rho)| \leq \frac{1}{2}|V_{p,j}(\rho)|.$$

Set

$$\zeta_p := \min_{j\in[J]} \zeta_{p,j}, \qquad a_p := \frac{1}{2}\min_{j\in[J]} |V_{p,j}(\rho)| > 0.$$

Then for every $j \in [J]$ and every p' with $\|p' - p\|_\infty \leq \zeta_p$,

$$|V_{p',j}(\rho)| \geq |V_{p,j}(\rho)| - |V_{p',j}(\rho) - V_{p,j}(\rho)| \geq a_p,$$

and the sign agrees with that of $V_{p,j}(\rho)$. □

Proposition I.4 (Continuity of the closed policywise testing root on a support-safe neighborhood)**.** *Assume $p \in \mathcal{P}_{\text{test}}^{1:J}$. Then there exists $\zeta_{\text{act}} > 0$ such that, for every $j \in [J]$, the map*

$$(w, p') \longmapsto \bar{\sigma}_j(w, p')$$

is continuous on

$$K_{\mu_{\text{nav}}} \times \mathcal{P}_{\zeta_{\text{act}}}(p), \qquad \mathcal{P}_{\zeta_{\text{act}}}(p) := \{p' \in \mathcal{P} : \|p' - p\|_\infty \leq \zeta_{\text{act}}\}.$$

Moreover, on this set,

$$\bar{\sigma}_j(w, p') = \sigma_{\text{NC}}^{(j)}(0, w, p').$$

Proof. Let $\zeta_{\mathrm{act}} := \zeta_p$, where ζ_p is given by Lemma I.3. Fix $j \in [J]$, and write

$$b_j := \mathrm{sign}(V_{p,j}(\rho)).$$

By Lemma I.3, every $p' \in \mathcal{P}_{\zeta_{\mathrm{act}}}(p)$ satisfies

$$\mathrm{sign}(V_{p',j}(\rho)) = b_j.$$

Hence the closed opposite-sign set can be written as the fixed compact set

$$\Gamma_j := \{q \in \mathcal{P} : \ b_j V_{q,j}(\rho) \le 0\}.$$

Indeed, for every $p' \in \mathcal{P}_{\zeta_{\mathrm{act}}}(p)$,

$$\{q \in \mathcal{P} : g_{p',j}(q) \le 0\} = \Gamma_j.$$

Therefore

$$\bar{\sigma}_j(w, p') = \min_{q \in \Gamma_j} \Phi(p'; w, q), \qquad (w, p') \in K_{\mu_{\mathrm{nav}}} \times \mathcal{P}_{\zeta_{\mathrm{act}}}(p).$$

By Appendix C, since every $p' \in \mathcal{P}_{\zeta_{\mathrm{act}}}(p)$ belongs to $\mathcal{P}^{1:J}_{\mathrm{test}}$ and every $w \in K_{\mu_{\mathrm{nav}}}$ has strictly positive coordinates, the closed and strict policywise testing roots agree:

$$\bar{\sigma}_j(w, p') = \sigma^{(j)}_{\mathrm{NC}}(0, w, p') \qquad \forall (w, p') \in K_{\mu_{\mathrm{nav}}} \times \mathcal{P}_{\zeta_{\mathrm{act}}}(p).$$

We use this equality below to obtain support-safe strict competitors.

We first prove lower semicontinuity. Let $(w_n, p_n) \to (w, p')$ with $w_n, w \in K_{\mu_{\mathrm{nav}}}$ and $p_n, p' \in \mathcal{P}_{\zeta_{\mathrm{act}}}(p)$. For each n, let $q_n \in \Gamma_j$ attain $\bar{\sigma}_j(w_n, p_n)$, i.e.

$$\bar{\sigma}_j(w_n, p_n) = \Phi(p_n; w_n, q_n).$$

By compactness of Γ_j, after passing to a subsequence we may assume

$$q_n \to q_\infty \in \Gamma_j.$$

Since Φ is lower semicontinuous in (p', w, q),

$$\bar{\sigma}_j(w, p') \le \Phi(p'; w, q_\infty) \le \liminf_{n \to \infty} \Phi(p_n; w_n, q_n) = \liminf_{n \to \infty} \bar{\sigma}_j(w_n, p_n).$$

Thus $\bar{\sigma}_j$ is lower semicontinuous.

We next prove upper semicontinuity. Fix

$$(w, p') \in K_{\mu_{\mathrm{nav}}} \times \mathcal{P}_{\zeta_{\mathrm{act}}}(p), \qquad (w_n, p_n) \to (w, p').$$

Let $\varepsilon > 0$. Since $\bar{\sigma}_j(w, p') = \sigma^{(j)}_{\mathrm{NC}}(0, w, p')$, there exists $q^{\mathrm{str}}_\varepsilon \in \mathcal{P}$ such that

$$g_{p',j}(q^{\mathrm{str}}_\varepsilon) < 0, \qquad \Phi(p'; w, q^{\mathrm{str}}_\varepsilon) \le \bar{\sigma}_j(w, p') + \varepsilon.$$

In particular, $q^{\mathrm{str}}_\varepsilon \in \Gamma_j$ with strict opposite sign. The displayed inequality implies that this competitor has finite $\Phi(p'; w, \cdot)$-value.

We now make the strict competitor rowwise full-support without leaving the strict opposite-sign branch. By Assumption 2.2, there exists a kernel q°_j with full row support and strict opposite sign:

$$b_j V_{q^\circ_j, j}(\rho) < 0.$$

For $\gamma \in (0, 1)$, define

$$q_{\varepsilon,\gamma} := (1 - \gamma) q^{\mathrm{str}}_\varepsilon + \gamma q^\circ_j.$$

Then $q_{\varepsilon,\gamma}$ has full row support. Since $q^{\mathrm{str}}_\varepsilon$ has strict opposite sign and $q \mapsto V_{q,j}(\rho)$ is continuous, $q_{\varepsilon,\gamma} \in \Gamma_j$ with strict opposite sign for all sufficiently small $\gamma > 0$. Moreover,

$$\Phi(p'; w, q_{\varepsilon,\gamma}) \to \Phi(p'; w, q^{\mathrm{str}}_\varepsilon) \qquad (\gamma \downarrow 0).$$

Indeed, $\Phi(p'; w, q_\varepsilon^{\mathrm{str}}) < \infty$ and $w \in K_{\mu_{\mathrm{nav}}}$, so $q_\varepsilon^{\mathrm{str}}$ dominates p' rowwise on every KL-relevant support. Since $q_{\varepsilon,\gamma} \to q_\varepsilon^{\mathrm{str}}$ rowwise and $q_{\varepsilon,\gamma} \geq (1-\gamma) q_\varepsilon^{\mathrm{str}}$, the rowwise KL terms converge.

Choose $\gamma = \gamma(\varepsilon) > 0$ small enough that

$$b_j V_{q_{\varepsilon,\gamma},j}(\rho) < 0$$

and

$$\Phi(p'; w, q_{\varepsilon,\gamma}) \leq \bar{\sigma}_j(w, p') + 2\varepsilon.$$

Set

$$\widetilde{q}_\varepsilon := q_{\varepsilon,\gamma}.$$

Then $\widetilde{q}_\varepsilon \in \Gamma_j$, has full row support, and satisfies

$$g_{p',j}(\widetilde{q}_\varepsilon) < 0, \qquad \Phi(p'; w, \widetilde{q}_\varepsilon) \leq \bar{\sigma}_j(w, p') + 2\varepsilon.$$

Since $(\tilde{p}, q) \mapsto g_{\tilde{p},j}(q)$ is continuous and $g_{p',j}(\widetilde{q}_\varepsilon) < 0$, there exists a neighborhood $\mathcal{V}_\varepsilon$ of p' such that

$$g_{\tilde{p},j}(\widetilde{q}_\varepsilon) < 0 \qquad \forall \tilde{p} \in \mathcal{V}_\varepsilon.$$

Thus $\widetilde{q}_\varepsilon$ remains strictly feasible for all nearby kernels. Since $\widetilde{q}_\varepsilon$ has full row support, the map

$$(\tilde{p}, \tilde{w}) \longmapsto \Phi(\tilde{p}; \tilde{w}, \widetilde{q}_\varepsilon)$$

is finite and continuous in a neighborhood of (p', w). Hence, for all sufficiently large n,

$$\Phi(p_n; w_n, \widetilde{q}_\varepsilon) \leq \Phi(p'; w, \widetilde{q}_\varepsilon) + \varepsilon \leq \bar{\sigma}_j(w, p') + 3\varepsilon.$$

Because $\widetilde{q}_\varepsilon$ is feasible for $\bar{\sigma}_j(w_n, p_n)$, we obtain

$$\bar{\sigma}_j(w_n, p_n) \leq \Phi(p_n; w_n, \widetilde{q}_\varepsilon) \leq \bar{\sigma}_j(w, p') + 3\varepsilon$$

for all sufficiently large n. Therefore

$$\limsup_{n\to\infty} \bar{\sigma}_j(w_n, p_n) \leq \bar{\sigma}_j(w, p').$$

Since $\varepsilon > 0$ is arbitrary, $\bar{\sigma}_j$ is upper semicontinuous. Together with lower semicontinuity, it is continuous on $K_{\mu_{\mathrm{nav}}} \times \mathcal{P}_{\zeta_{\mathrm{act}}}(p)$. □

For $p' \in \mathcal{P}_{\zeta_{\mathrm{act}}}(p)$ and $w \in K_{\mu_{\mathrm{nav}}}$, define

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

This agrees with the policywise sign-flipped frontier used elsewhere in the paper on the local testing neighborhood.

Proposition I.5 (Uniform local policywise testing separation below the frontier)**.** *Assume $p \in \mathcal{P}_{\mathrm{test}}^{1:J}$. Fix $\eta > 0$ and $\bar{\sigma} < \infty$. Then there exist constants*

$$\zeta_{\mathrm{sep}}(\eta, \bar{\sigma}) > 0, \qquad m_{\mathrm{sep}}(\eta, \bar{\sigma}) > 0$$

such that the following holds.

For every policy index $j \in [J]$, every

$$p' \in \mathcal{P}, \qquad \|p' - p\|_\infty \leq \zeta_{\mathrm{sep}}(\eta, \bar{\sigma}), \qquad w \in K_{\mu_{\mathrm{nav}}}, \qquad \sigma \in [0, \bar{\sigma}],$$

if

$$\sigma \leq F_{p'}^{(j)}(w) - \eta,$$

then

$$u_{\mathrm{NO}}^{(j)}(\sigma, w, p') \geq m_{\mathrm{sep}}(\eta, \bar{\sigma}).$$

Proof. We argue by contradiction. Assume the claim is false. Then for every $n \geq 1$ there exist a policy index $j_n \in [J]$ and

$$p_n \in \mathcal{P}, \qquad \|p_n - p\|_\infty \leq \frac{1}{n}, \qquad w_n \in K_{\mu_{\text{nav}}}, \qquad \sigma_n \in [0, \bar{\sigma}],$$

such that

$$\sigma_n \leq F_{p_n}^{(j_n)}(w_n) - \eta, \qquad u_{\text{NO}}^{(j_n)}(\sigma_n, w_n, p_n) \leq \frac{1}{n}.$$

Since J is finite, after passing to a subsequence we may assume that $j_n \equiv j_\infty$ is constant. By compactness of $K_{\mu_{\text{nav}}}$, $[0, \bar{\sigma}]$, and a small local neighborhood of p, after passing to a further subsequence we may assume

$$p_n \to p_\infty, \qquad w_n \to w_\infty, \qquad \sigma_n \to \sigma_\infty,$$

with

$$p_\infty = p, \qquad w_\infty \in K_{\mu_{\text{nav}}}, \qquad \sigma_\infty \in [0, \bar{\sigma}].$$

For all sufficiently large n, $p_n \in \mathcal{P}_{\zeta_{\text{act}}}(p)$.

For each n, choose $q_n \in \mathcal{P}$ attaining $u_{\text{NO}}^{(j_\infty)}(\sigma_n, w_n, p_n)$, i.e.

$$\Phi(p_n; w_n, q_n) \leq \sigma_n, \qquad g_{p_n, j_\infty}(q_n) = u_{\text{NO}}^{(j_\infty)}(\sigma_n, w_n, p_n) \leq \frac{1}{n}.$$

By compactness of $\mathcal{P}$, after passing to a further subsequence,

$$q_n \to q_\infty \in \mathcal{P}.$$

Since Φ is lower semicontinuous,

$$\Phi(p_\infty; w_\infty, q_\infty) \leq \liminf_{n\to\infty} \Phi(p_n; w_n, q_n) \leq \lim_{n\to\infty} \sigma_n = \sigma_\infty.$$

Since $g_{p', j_\infty}(q)$ is continuous in (p', q),

$$g_{p_\infty, j_\infty}(q_\infty) = \lim_{n\to\infty} g_{p_n, j_\infty}(q_n) \leq 0.$$

Hence q_∞ is feasible for the closed testing root problem at (w_∞, p_∞) for policy j_∞. Therefore

$$\bar{\sigma}_{j_\infty}(w_\infty, p_\infty) \leq \sigma_\infty. \tag{16}$$

For all sufficiently large n, Proposition I.4 and the local definition of $F^{(j)}$ give

$$F_{p_n}^{(j_\infty)}(w_n) = \bar{\sigma}_{j_\infty}(w_n, p_n).$$

Since

$$\sigma_n \leq F_{p_n}^{(j_\infty)}(w_n) - \eta,$$

we obtain

$$\sigma_n \leq \bar{\sigma}_{j_\infty}(w_n, p_n) - \eta.$$

Passing to the limit and using continuity of $\bar{\sigma}_{j_\infty}$,

$$\sigma_\infty \leq \bar{\sigma}_{j_\infty}(w_\infty, p_\infty) - \eta. \tag{17}$$

Combining (16) and (17) yields

$$\bar{\sigma}_{j_\infty}(w_\infty, p_\infty) \leq \sigma_\infty \leq \bar{\sigma}_{j_\infty}(w_\infty, p_\infty) - \eta,$$

a contradiction.

Therefore the claimed constants

$$\zeta_{\text{sep}}(\eta, \bar{\sigma}) > 0, \qquad m_{\text{sep}}(\eta, \bar{\sigma}) > 0$$

exist. □

J. Proof of Theorem 3.7

In this appendix we prove Theorem 3.7. The argument combines the three layers of PT-ACE(μ) developed in the previous appendices.

Appendix G proves that the exchange layer learns an allocation whose bottleneck frontier value is close to the simultaneous lower-bound benchmark on a compact interior slice. Appendix H transfers this optimization output to the realized trajectory and gives a lower bound on the empirical allocation $\hat{w}(T)$. Appendix I turns this frontier lower bound into a stopping statement for the policywise testing frontiers $u_{\mathrm{NO}}^{(j)}$, uniformly on the compact slice $K_{\mu_{\mathrm{nav}}}$.

Throughout this appendix,

$$\mathcal{C}_T(\zeta) := \mathcal{C}_T^{\mathrm{ker}}(\zeta) \cap \mathcal{C}_T^{\mathrm{nav}}(\zeta).$$

Here $\mathcal{C}_T^{\mathrm{nav}}(\zeta)$ is the navigation event of Appendix H, defined using the frozen navigation policies (χ_t) and the tail window

$$\vartheta_T := \lceil T^{3/4} \rceil.$$

We split the proof into four steps. First, Subsection J.1 proves the δ-probably correct guarantee from the policywise certified testing brackets returned by the stopping oracle. Second, Subsection J.2 establishes a deterministic implication on the good event: once T is larger than a suitable multiple of $\log(1/\delta)$, the queried budget σ_T lies below every policywise testing frontier by a fixed margin, and the algorithm must stop by time T. Third, Subsection J.3 combines this deterministic implication with the kernel and navigation probability bounds to derive a tail bound on τ_δ. Finally, Subsection J.4 integrates the tail bound to obtain the asymptotic upper bound on $\mathbb{E}_p[\tau_\delta]$.

J.1. The δ-probably correct guarantee

We first prove that the stopping rule of Algorithm 1 is δ-probably correct for simultaneous testing.

Let

$$\beta(t,\delta) := \log(1/\delta) + (|\mathcal{S}|-1)\sum_{s,a}\log\Big(e\big[1+N_{sa}(t)/(|\mathcal{S}|-1)\big]\Big), \qquad \sigma_t := \frac{\beta(t,\delta)}{t \vee 1}.$$

We use the self-normalized KL guarantee: for every $p \in \mathcal{P}$,

$$\mathbb{P}_p\left(\exists t \geq 1 : \sum_{s,a} N_{sa}(t)\mathrm{KL}_{sa}(\hat{p}_t, p) \geq \beta(t,\delta)\right) \leq \delta. \tag{18}$$

Define

$$\mathcal{E}_\delta := \left\{\forall t \geq 1,\ \sum_{s,a} N_{sa}(t)\mathrm{KL}_{sa}(\hat{p}_t, p) < \beta(t,\delta)\right\}.$$

By (18),

$$\mathbb{P}_p(\mathcal{E}_\delta) \geq 1-\delta.$$

Now fix $t \geq 1$. On $\mathcal{E}_\delta$,

$$\sum_{s,a} N_{sa}(t)\mathrm{KL}_{sa}(\hat{p}_t, p) < \beta(t,\delta),$$

and hence

$$\Phi(\hat{p}_t; \hat{w}(t), p) < \sigma_t.$$

Thus, on $\mathcal{E}_\delta$, the true kernel p belongs to the weighted-KL plausible set queried by every policywise stopping oracle at every round t.

Let τ_δ be the stopping time of Algorithm 1. Suppose that the algorithm stops at time τ_δ. By the stopping rule,

$$L_{\tau_\delta,j}^{\mathrm{stop}} > 0 \qquad \forall j \in [J],$$

and the returned answer is

$$\hat{\imath}_{\tau_\delta}^{\mathrm{stop}} := \big(\mathrm{Ans}_1(\hat{p}_{\tau_\delta}), \ldots, \mathrm{Ans}_J(\hat{p}_{\tau_\delta})\big).$$

By the certified policywise stopping-oracle contract stated in Proposition 3.4, for each $j \in [J]$,

$$L^{\mathrm{stop}}_{\tau_\delta,j} \le u^{(j)}_{\mathrm{NO}}\big(\sigma_{\tau_\delta}; \hat{w}(\tau_\delta), \hat{p}_{\tau_\delta}\big) \le U^{\mathrm{stop}}_{\tau_\delta,j}.$$

Therefore

$$u^{(j)}_{\mathrm{NO}}\big(\sigma_{\tau_\delta}; \hat{w}(\tau_\delta), \hat{p}_{\tau_\delta}\big) > 0 \qquad \forall j \in [J].$$

By definition of $u^{(j)}_{\mathrm{NO}}$, every kernel $q \in \mathcal{P}$ such that

$$\Phi(\hat{p}_{\tau_\delta}; \hat{w}(\tau_\delta), q) \le \sigma_{\tau_\delta}$$

satisfies

$$V_{\hat{p}_{\tau_\delta},j}(\rho) V_{q,j}(\rho) > 0.$$

On $\mathcal{E}_\delta$, the true kernel p is feasible for this plausible set. Therefore

$$V_{\hat{p}_{\tau_\delta},j}(\rho) V_{p,j}(\rho) > 0 \qquad \forall j \in [J],$$

and hence

$$\mathrm{Ans}_j(\hat{p}_{\tau_\delta}) = \mathrm{Ans}_j(p) \qquad \forall j \in [J].$$

Thus, on $\mathcal{E}_\delta$,

$$\hat{\imath}^{\mathrm{stop}}_{\tau_\delta} = \mathrm{Ans}_{1:J}(p).$$

Consequently,

$$\mathbb{P}_p\big(\hat{\imath}^{\mathrm{stop}}_{\tau_\delta} \ne \mathrm{Ans}_{1:J}(p)\big) \le \mathbb{P}_p(\mathcal{E}^c_\delta) \le \delta.$$

Therefore PT-ACE(μ) is δ-probably correct.

It remains to prove the asymptotic upper bound on $\mathbb{E}_p[\tau_\delta]$.

J.2. A deterministic stopping implication on the good event

Define

$$\Lambda^{\mathrm{test}}_\mu(p) := (1 + c(\mu, p))\, T^\star(p), \qquad a^{\mathrm{test}}_\mu(p) := \frac{1}{\Lambda^{\mathrm{test}}_\mu(p)} = \frac{1}{(1 + c(\mu, p))\, T^\star(p)}.$$

Equivalently,

$$a^{\mathrm{test}}_\mu(p) = \frac{1}{1 + c(\mu, p)} \max_{w \in \Sigma_p} F^{1:J}_p(w).$$

This is the frontier level that drives the stopping-time upper bound.

Let

$$\underline{\alpha}_{\mathrm{nav}}(p, \mu) > 0$$

be the constant from Lemma H.8, and define

$$\mu_\star := \frac{1}{2} \min\{\mu, \underline{\alpha}_{\mathrm{nav}}(p, \mu)\}.$$

Let $\psi_{p,\mu_\star}$ denote a continuity modulus of $F^{1:J}_p$ on the compact slice $K_{\mu_\star}$, as in Proposition H.9.

For later use, define the deterministic stopping-side remainder

$$r_{\mathrm{stop}}(T, \mu, \zeta) := r^{\mathrm{tail}}_{\mathrm{ex}}(T, \mu, \zeta) + \psi_{p,\mu_\star}\Big(\zeta + C_\mu(p)\, r_{\mathrm{flow}}(T, \zeta)\Big).$$

By Proposition G.16, Proposition G.17, and Proposition H.9,

$$\lim_{\zeta \downarrow 0} \limsup_{T \to \infty} r_{\mathrm{stop}}(T, \mu, \zeta) = 0.$$

We first record the lower bound on the true bottleneck frontier value obtained from navigation.

Lemma J.1 (Navigation-side lower bound on the simultaneous testing frontier)**.** *Fix $0 < \zeta \le \underline{\alpha}_{\mathrm{nav}}(p,\mu)/2$. Then, on $\mathcal{C}_T(\zeta)$, for all sufficiently large T,*

$$F_p^{1:J}(\hat{w}(T)) \ge a_\mu^{\mathrm{test}}(p) - r_{\mathrm{stop}}(T,\mu,\zeta).$$

Proof. This is exactly Proposition H.9, rewritten using the definition of $a_\mu^{\mathrm{test}}(p)$ and $r_{\mathrm{stop}}(T,\mu,\zeta)$. □

We next transfer this lower bound from $F_p^{1:J}$ to $F_{\hat{p}_T}^{1:J}$ on the compact interior slice.

Lemma J.2 (Kernel perturbation of the simultaneous testing frontier on $K_{\mu_\star}$)**.** *There exist $\zeta_\mu^\circ > 0$ and a deterministic modulus $\omega_\mu : [0,\zeta_\mu^\circ] \to \mathbb{R}_+$ with*

$$\omega_\mu(r) \to 0 \qquad \text{as } r \downarrow 0,$$

such that for every $0 < \zeta \le \zeta_\mu^\circ$,

$$\sup_{\substack{w \in K_{\mu_\star} \\ p' \in \mathcal{P}:\ \|p'-p\|_\infty \le \zeta}} \left| F_{p'}^{1:J}(w) - F_p^{1:J}(w) \right| \le \omega_\mu(\zeta).$$

Proof. By Lemma G.12, applied on the compact interior slice $K_{\mu_\star}$, the map

$$(p', w) \longmapsto F_{p'}^{1:J}(w)$$

is continuous on a neighborhood of p times $K_{\mu_\star}$. Compactness gives uniform continuity, and the claimed modulus is the resulting local modulus around p. □

Lemma J.3 (Deterministic stopping on the good event)**.** *Fix $\lambda > \Lambda_\mu^{\mathrm{test}}(p)$. Then there exist constants*

$$\zeta_\lambda > 0, \qquad \eta_\lambda > 0, \qquad m_\lambda > 0, \qquad T_\lambda < \infty, \qquad \delta_\lambda > 0$$

such that for every $T \ge T_\lambda$ and every $0 < \delta \le \delta_\lambda$, the following implication holds on $\mathcal{C}_T(\zeta_\lambda)$:

$$T \ge \lambda \log(1/\delta) \quad \Longrightarrow \quad \tau_\delta \le T.$$

Proof. Fix $\lambda > \Lambda_\mu^{\mathrm{test}}(p)$. Then

$$\frac{1}{\lambda} < a_\mu^{\mathrm{test}}(p).$$

Choose $\eta_\lambda > 0$ such that

$$\frac{1}{\lambda} + 3\eta_\lambda < a_\mu^{\mathrm{test}}(p).$$

Choose $\bar{\sigma} < \infty$ large enough so that, for all sufficiently small δ, every queried budget

$$\sigma_T = \frac{\beta(T,\delta)}{T}, \qquad T \ge \lambda \log(1/\delta),$$

belongs to $[0,\bar{\sigma}]$.

By Proposition I.5, applied with $\eta = \eta_\lambda$ and this $\bar{\sigma}$, there exist

$$\zeta_{\mathrm{sep}}(\eta_\lambda,\bar{\sigma}) > 0, \qquad m_{\mathrm{sep}}(\eta_\lambda,\bar{\sigma}) > 0.$$

Set

$$m_\lambda := m_{\mathrm{sep}}(\eta_\lambda,\bar{\sigma}).$$

Since

$$\omega_\mu(\zeta) \to 0 \quad (\zeta \downarrow 0), \qquad \lim_{\zeta \downarrow 0} \limsup_{T \to \infty} r_{\mathrm{stop}}(T,\mu,\zeta) = 0,$$

and since $\xi_T^{\mathrm{stop}} \to 0$ by the tolerance condition (4), we may choose $\zeta_\lambda > 0$ small enough so that

$$\zeta_\lambda \le \min\Big\{\zeta_\mu^\circ,\ \zeta_{\mathrm{sep}}(\eta_\lambda,\bar{\sigma}),\ \zeta_p,\ \underline{\alpha}_{\mathrm{nav}}(p,\mu)/2\Big\},$$

and then choose $T_\lambda < \infty$ so large that for all $T \geq T_\lambda$,

$$r_{\text{stop}}(T, \mu, \zeta_\lambda) + \omega_\mu(\zeta_\lambda) \leq \eta_\lambda, \qquad \xi_T^{\text{stop}} \leq \frac{m_\lambda}{2}. \tag{19}$$

Here ζ_p is the local sign-stability radius from Lemma I.3.

Now work on $\mathcal{C}_T(\zeta_\lambda)$, and suppose $T \geq T_\lambda$. By Corollary I.2 and Proposition H.9, after increasing T_λ if necessary, the empirical allocation $\hat{w}(T)$ belongs to both $K_{\mu_{\text{nav}}}$ and $K_{\mu_\star}$. Also,

$$\|\hat{p}_T - p\|_\infty \leq \zeta_\lambda.$$

By Lemma J.1,

$$F_p^{1:J}(\hat{w}(T)) \geq a_\mu^{\text{test}}(p) - r_{\text{stop}}(T, \mu, \zeta_\lambda).$$

Using Lemma J.2, we obtain

$$F_{\hat{p}_T}^{1:J}(\hat{w}(T)) \geq F_p^{1:J}(\hat{w}(T)) - \omega_\mu(\zeta_\lambda).$$

By (19), this yields

$$F_{\hat{p}_T}^{1:J}(\hat{w}(T)) \geq a_\mu^{\text{test}}(p) - \eta_\lambda.$$

Next, if $T \geq \lambda \log(1/\delta)$, then by the definition of $\beta(T, \delta)$,

$$\sigma_T = \frac{\beta(T, \delta)}{T} \leq \frac{1}{\lambda} + \frac{|\mathcal{S}||\mathcal{A}|(|\mathcal{S}| - 1)\log\big(e[1 + T/(|\mathcal{S}| - 1)]\big)}{T}.$$

The second term tends to 0 uniformly over $T \geq \lambda \log(1/\delta)$ as $\delta \downarrow 0$. Therefore, after shrinking $\delta_\lambda > 0$ if necessary, for every $0 < \delta \leq \delta_\lambda$ and every $T \geq \lambda \log(1/\delta)$,

$$\sigma_T \leq \frac{1}{\lambda} + \eta_\lambda. \tag{20}$$

Combining the last two displays and the choice of η_λ, we get

$$\sigma_T \leq F_{\hat{p}_T}^{1:J}(\hat{w}(T)) - \eta_\lambda.$$

Since

$$F_{\hat{p}_T}^{1:J}(\hat{w}(T)) = \min_{j \in [J]} F_{\hat{p}_T}^{(j)}(\hat{w}(T)),$$

it follows that for every $j \in [J]$,

$$\sigma_T \leq F_{\hat{p}_T}^{(j)}(\hat{w}(T)) - \eta_\lambda.$$

The algorithm queries the stopping oracle at time T, unless it has already stopped earlier. Indeed, on the good event and for $T \geq T_\lambda$, every state-action pair has positive empirical count and $\hat{p}_T \in \mathcal{P}_{\text{test}}^{1:J}$, by the empirical-interiority and local sign-stability statements above. If the algorithm has already stopped, then $\tau_\delta \leq T$ is immediate. Otherwise, it queries every policywise oracle at time T.

Since $\sigma_T \in [0, \bar{\sigma}]$, $\|\hat{p}_T - p\|_\infty \leq \zeta_{\text{sep}}(\eta_\lambda, \bar{\sigma})$, and $\hat{w}(T) \in K_{\mu_{\text{nav}}}$, Proposition I.5 applies to each j and gives

$$u_{\text{NO}}^{(j)}(\sigma_T, \hat{w}(T), \hat{p}_T) \geq m_\lambda \qquad \forall j \in [J].$$

By the certified policywise stopping-oracle contract stated in Proposition 3.4, the policywise stopping oracles return certified brackets

$$L_{T,j}^{\text{stop}} \leq u_{\text{NO}}^{(j)}(\sigma_T, \hat{w}(T), \hat{p}_T) \leq U_{T,j}^{\text{stop}}, \qquad U_{T,j}^{\text{stop}} - L_{T,j}^{\text{stop}} \leq \xi_T^{\text{stop}}.$$

Using (19),

$$L_{T,j}^{\text{stop}} \geq u_{\text{NO}}^{(j)}(\sigma_T, \hat{w}(T), \hat{p}_T) - \xi_T^{\text{stop}} \geq m_\lambda - \frac{m_\lambda}{2} = \frac{m_\lambda}{2} > 0 \qquad \forall j \in [J].$$

Hence the stopping condition of Algorithm 1 is satisfied at time T. Therefore

$$\tau_\delta \leq T.$$

This proves the deterministic implication. □

J.3. Tail bound for the stopping time

Lemma J.4 (Tail bound for τ_δ)**.** *Fix $\lambda > \Lambda_\mu^{\mathrm{test}}(p)$. Then there exist constants*

$$\zeta_\lambda > 0, \qquad \delta_\lambda > 0, \qquad B_\lambda > 0, \qquad C_\lambda > 0,$$

such that for every $0 < \delta \le \delta_\lambda$, if

$$n_\delta := \Big\lceil \lambda \log(1/\delta) \Big\rceil,$$

then for every integer $m \ge 1$,

$$\mathbb{P}_p(\tau_\delta > mn_\delta) \le B_\lambda(mn_\delta)\exp\big(-C_\lambda(mn_\delta)^{1/4}\big) + B_\lambda \exp\big(-C_\lambda mn_\delta\big).$$

Proof. Fix $\lambda > \Lambda_\mu^{\mathrm{test}}(p)$, and let $\zeta_\lambda, \delta_\lambda$ be given by Lemma J.3. Define

$$T_m := mn_\delta, \qquad m \ge 1.$$

Since

$$n_\delta = \Big\lceil \lambda \log(1/\delta) \Big\rceil,$$

we have

$$T_m \ge n_\delta \ge \lambda \log(1/\delta).$$

Therefore, for every $m \ge 1$, Lemma J.3 implies

$$\{\tau_\delta > T_m\} \subset \big(\mathcal{C}_{T_m}^{\mathrm{ker}}(\zeta_\lambda) \cap \mathcal{C}_{T_m}^{\mathrm{nav}}(\zeta_\lambda)\big)^c.$$

Hence

$$\mathbb{P}_p(\tau_\delta > T_m) \le \mathbb{P}_p\big((\mathcal{C}_{T_m}^{\mathrm{ker}}(\zeta_\lambda))^c\big) + \mathbb{P}_p\big((\mathcal{C}_{T_m}^{\mathrm{nav}}(\zeta_\lambda))^c\big).$$

Apply Proposition F.7 to the first term and Proposition H.5 to the second one. This yields constants $B_\lambda, C_\lambda > 0$, depending only on (λ, μ, p), such that for every $m \ge 1$ and every sufficiently small δ,

$$\mathbb{P}_p\big((\mathcal{C}_{T_m}^{\mathrm{ker}}(\zeta_\lambda))^c\big) \le B_\lambda T_m \exp\big(-C_\lambda T_m^{1/4}\big),$$

and

$$\mathbb{P}_p\big((\mathcal{C}_{T_m}^{\mathrm{nav}}(\zeta_\lambda))^c\big) \le B_\lambda \exp\big(-C_\lambda T_m\big).$$

Substituting $T_m = mn_\delta$ proves the claim. □

J.4. Expectation upper bound

Proof of Theorem 3.7. The δ-probably correct statement was established in Subsection J.1. We now prove the asymptotic sample-complexity bound.

Fix $\lambda > \Lambda_\mu^{\mathrm{test}}(p)$, and define

$$n_\delta := \Big\lceil \lambda \log(1/\delta) \Big\rceil.$$

Since τ_δ is a nonnegative integer-valued random variable, the tail-sum identity gives

$$\mathbb{E}_p[\tau_\delta] = \sum_{k \ge 0} \mathbb{P}_p(\tau_\delta > k).$$

Grouping the sum into blocks of length n_δ, we obtain

$$\mathbb{E}_p[\tau_\delta] \le n_\delta \sum_{m \ge 0} \mathbb{P}_p(\tau_\delta > mn_\delta).$$

The term $m = 0$ equals 1. For $m \ge 1$, Lemma J.4 yields

$$\mathbb{P}_p(\tau_\delta > mn_\delta) \le B_\lambda(mn_\delta)\exp\big(-C_\lambda(mn_\delta)^{1/4}\big) + B_\lambda \exp\big(-C_\lambda mn_\delta\big).$$

Therefore

$$\mathbb{E}_p[\tau_\delta] \le n_\delta \left(1 + \sum_{m\ge 1} B_\lambda(mn_\delta) \exp\bigl(-C_\lambda (mn_\delta)^{1/4}\bigr) + \sum_{m\ge 1} B_\lambda \exp\bigl(-C_\lambda m n_\delta\bigr) \right).$$

Both series in parentheses converge to 0 as $\delta \downarrow 0$. Indeed, $n_\delta \to \infty$, and for the first series, for all sufficiently small δ,

$$(mn_\delta) \exp\bigl(-C_\lambda (mn_\delta)^{1/4}\bigr) \le m\, e^{-\frac{C_\lambda}{2} m^{1/4}},$$

because $n_\delta e^{-(C_\lambda/2) n_\delta^{1/4}} \le 1$ eventually. The right-hand side is summable over $m \ge 1$. For the second series,

$$\exp\bigl(-C_\lambda m n_\delta\bigr) \le e^{-C_\lambda m},$$

which is also summable over $m \ge 1$. Hence dominated convergence gives

$$\sum_{m\ge 1} B_\lambda(mn_\delta) \exp\bigl(-C_\lambda (mn_\delta)^{1/4}\bigr) \to 0,$$

and

$$\sum_{m\ge 1} B_\lambda \exp\bigl(-C_\lambda m n_\delta\bigr) \to 0.$$

We conclude that

$$\mathbb{E}_p[\tau_\delta] \le n_\delta (1 + o(1)) = \lambda \log(1/\delta)\,(1 + o(1)).$$

Therefore

$$\limsup_{\delta\downarrow 0} \frac{\mathbb{E}_p[\tau_\delta]}{\log(1/\delta)} \le \lambda.$$

Since $\lambda > \Lambda_\mu^{\mathrm{test}}(p)$ was arbitrary,

$$\limsup_{\delta\downarrow 0} \frac{\mathbb{E}_p[\tau_\delta]}{\log(1/\delta)} \le \Lambda_\mu^{\mathrm{test}}(p).$$

Finally, by definition,

$$\Lambda_\mu^{\mathrm{test}}(p) = (1 + c(\mu, p))\, T^\star(p),$$

where $c(\mu, p)$ is the constant from Lemma G.1. Thus

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

Moreover,

$$c(\mu, p) = \frac{\theta_\mu}{1 - \theta_\mu}, \qquad \theta_\mu = \frac{2\mu}{\underline{\alpha}_p},$$

for

$$0 < \mu \le \min\Bigl\{ \frac{1}{2|\mathcal{S}||\mathcal{A}|}, \frac{\underline{\alpha}_p}{4} \Bigr\}.$$

Hence $c(\mu, p) \to 0$ as $\mu \downarrow 0$. This completes the proof. □

K. 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 $\{\pi^{(j)}\}_{j=1}^{J}$
$V_{p,j}(\rho)$	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	fixed-floor slice $\{w : w_{sa} \geq \mu \, \forall (s,a)\}$
$\widetilde{\Sigma}_{p'}^{\mu}$	fixed-floor occupancy domain $\Sigma_{p'} \cap K_\mu$
$\Phi(p; \omega, q)$	weighted rowwise KL information $\sum_{s,a} \omega_{sa} \mathrm{KL}_{sa}(p,q)$
$F_p^{(j)}, F_p^{1:J}$	policywise and joint bottleneck frontiers
$T^\star(p)$	characteristic time, $1/T^\star(p) = \max_{\omega \in \Sigma_p} F_p^{1:J}(\omega)$
$x_t, y_{t+1}, \bar{x}_t$	exchange center, predictor, and averaged corrected center
χ_t	frozen navigation policy induced by $\bar{x}_t$
g_t, q_t, j_t	cut vector, witness kernel, and bottleneck 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	tail-window indices $\lceil T^{1/4} \rceil$, $\lceil T^{3/4} \rceil$
τ_δ	stopping time of PT-ACE(μ)

Table 1. Frequently used notation.