WHEN OCCUPANCY MULTIPLICITY MAKES AUDITING HARD: A CONDITIONAL QUERY LOWER BOUND

Anonymous authors
Paper under double-blind review

ABSTRACT

Near-optimal reinforcement-learning policies can have different reachable behavior, but their number alone does not determine how difficult they are to audit. We define occupancy Rashomon capacity as the metric entropy of near-optimal discounted state-action occupancies within a declared deployment class. A separated family of M policies with b-sparse exact local signatures requires $\Omega(M/b)$ state queries for occupancy identification. Expressing this bound exponentially in capacity requires the family to realize the cover number; it is the same search bound in different units. A finite hidden-branch MDP attains the bound, whereas a shortest-path family has high capacity and admits efficient path-following audits. We derive a separate root-reset episode law, a noisy-trigger lower bound of $\Omega(M/\beta)$ for per-sample KL signal β, and reference-policy KL conditions for occupancy-capacity collapse. Learned SAC and DQN families provide finite, pipeline-relative diagnostics of occupancy separation and local agreement. These diagnostics distinguish certified separation from probe-dependent tolerance sparsity and do not establish uniform sparsity or exponential auditing hardness in deployed systems.

1 INTRODUCTION

Policies with similar reward can behave differently at bottlenecks, on rare observations, or after a trigger. An auditor must therefore determine which behavior has been deployed, even when return evaluation does not distinguish the candidates. This raises a structural question: when does a large family of near-optimal policies make auditing difficult, and when can a few informative observations identify the deployed behavior?

The Rashomon effect describes the multiplicity of predictive models with similar loss (Breiman, 2001; Semenova et al., 2022); Rashomon Capacity quantifies predictive multiplicity for classifiers (Hsu & Calmon, 2022). In sequential decision-making, policies also determine which observations become reachable. We work with discounted state-action occupancies so that policy differences on unreachable states do not create artificial multiplicity. Our capacity is the logarithm of a covering number at a specified occupancy resolution, computed within the deployment class the auditor actually considers.

The main result separates multiplicity from information. A family may contain exponentially many occupancy classes and still admit a short audit if local queries partition the candidates into balanced groups. Hardness arises under an additional condition: at every query state, all but at most b candidates share one exact response. For an occupancy-separated family of size M, this condition bounds identification success after n queries by $(nb + 1)/M$. A hidden-branch MDP realizes all requirements simultaneously: feasible discounted flow, realizable policies, optimal return, occupancy separation, and sparse action signatures. A shortest-path grid supplies the converse lesson: many optimal occupancies remain easy to identify when successive actions reveal the route.

The $\Omega(M/b)$ counting argument is a hidden-index search bound. Rewriting it as an exponential in occupancy capacity does not strengthen it. The RL-specific issue is whether a deployment family realizes both the occupancy geometry and the local information restriction, and whether the auditor can reach informative states. We address this issue through a precise deployment-class construction, a separate root-reset analysis, and learned-family diagnostics that retain uncertainty about return eligibility and the limitations of finite probes.

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

We also distinguish exact identification from noisy detection and mitigation. A hidden anomalous trigger requires $\Omega(M/\beta)$ samples when each observation has KL signal β; an extra $\log M$ factor is not universal. A forward policy-KL penalty controls occupancy distance to a reference under explicit support and optimality assumptions. Its worst-case complete-collapse threshold is much stronger than the penalties used in the empirical study. Full proofs, the static recognition bound, oracle-cover verification, and all additional benchmarks appear in the appendix.

2 Related Work

Rashomon sets, model reliance, predictive multiplicity, and interpretability study how comparable predictive performance can coexist with different models or explanations (Fisher et al., 2019; Marx et al., 2020; Semenova et al., 2022; Müller et al., 2023; Rudin et al., 2022). The capacity of Hsu & Calmon (2022) concerns probabilistic classification. We instead measure discounted occupancy geometry and combine it with an explicit oracle restriction. Neither a large covering number nor a large candidate count alone implies our auditing lower bound.

Policy-diversity methods such as DIAYN (Eysenbach et al., 2018) and structured maximum-entropy RL (Kumar et al., 2020) generate distinct behaviors. Reward misspecification and reward gaming explain why performance objectives can leave behavioral choices unconstrained (Pan et al., 2022; Skalse et al., 2022; Casper et al., 2023). Backdoor and poisoning studies demonstrate concrete vulnerabilities in learned policies (Kiourti et al., 2020; Wang et al., 2021; Ashcraft & Karra, 2021; Rakhsha et al., 2020). These studies motivate auditing; they do not by themselves establish sparse signatures for an occupancy-separated deployment family. Formal verification and logically constrained RL use stronger structural access to obtain safety guarantees (Fulton & Platzer, 2018; Hasanbeig et al., 2020).

Our proof tools are classical: Yao's principle (Yao, 1977), adaptive change of measure, KL convexity, and Le Cam/Pinsker inequalities (LeCam, 1973; Yu, 1997; Tsybakov, 2009). The static subset-recognition bound is an elementary information argument in the cell-probe spirit (Yao, 1981). We use these tools to identify the assumptions under which occupancy multiplicity obstructs a policy audit, without claiming a new general search lower bound or a refined time–space tradeoff.

3 Occupancy Auditing and the Deployment Class

A finite discounted MDP is $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma, \mu_0)$, with finite state and action sets, rewards in $[0,1]$, and $\gamma \in [0,1)$. Write $H = (1-\gamma)^{-1}$. A stationary policy π induces

$$d^\pi(s,a) = (1-\gamma)\sum_{t\geq 0}\gamma^t \Pr^\pi(s_t = s, a_t = a), \qquad V^\pi = \sum_{s,a} d^\pi(s,a) r(s,a). \tag{1}$$

Thus d^π is a probability distribution and $V^\pi \in [0,1]$ is normalized discounted return. Feasibility requires $\sum_a d(s,a) = (1-\gamma)\mu_0(s) + \gamma \sum_{s',a'} P(s \mid s',a') d(s',a')$. Let $V^\star = \max_\pi V^\pi$ be the global optimum. For a declared policy class $\mathcal{F}$ and tolerance $\varepsilon \geq 0$, set

$$\mathfrak{R}_\varepsilon(\mathcal{M};\mathcal{F}) = \{\pi \in \mathcal{F} : V^\star - V^\pi \leq \varepsilon\}, \qquad \mathcal{D}_\varepsilon(\mathcal{M};\mathcal{F}) = \{d^\pi : \pi \in \mathfrak{R}_\varepsilon(\mathcal{M};\mathcal{F})\}.$$

Definition 1 (Deployment-class occupancy Rashomon capacity)**.** *For a nonempty near-optimal occupancy set,*

$$H^{\mathcal{F}}_{\mathrm{opt}}(\varepsilon) = \log_2 \mathcal{N}\big(\mathcal{D}_\varepsilon(\mathcal{M};\mathcal{F}), 1/H, \mathrm{TV}\big), \tag{2}$$

where $\mathcal{N}$ is the minimum number of closed TV balls needed to cover the set. Centers may lie in the ambient probability simplex. An empty eligible class is reported as empty rather than assigned a finite logarithmic capacity.

Definition 2 (Occupancy-class auditing)**.** *Given oracle access to $\pi \in \mathfrak{R}_\varepsilon(\mathcal{M};\mathcal{F})$, an auditor succeeds at scale $1/H$ if it returns $\widehat{d}$ with $\mathrm{TV}(\widehat{d}, d^\pi) \leq 1/H$ with probability at least $2/3$.*

This target identifies reachable state-action behavior, not the full policy function. Exact local queries return $\pi(\cdot \mid s)$ at a chosen state; sample local queries return one action drawn from it. An exact response can simulate a sample response, so an exact-query lower bound also applies to sample queries. Episode access has a different unit of cost and requires a separate argument.

3.1 Declaring and evaluating a finite pipeline family

An operational class is $\mathcal{F}_{\mathcal{P}} = \{\pi_1, \dots, \pi_K\}$, the outputs admitted by a pipeline manifest. The manifest fixes the environment and wrappers, algorithm, training budgets, seeds, checkpoint-selection rule, checkpoint hashes, preprocessing, and deployed action rule. It identifies the candidates to which a conclusion applies. It cannot certify that an incompletely known pipeline has no other outputs.

Global near-optimality remains relative to $V^\star$. When no valid global upper certificate is available, learned-family results instead use $V^\star_{\mathcal{P}} = \max_i V^{\pi_i}$. Suppose held-out evaluation gives simultaneous intervals $[L_i, U_i]$ for all K returns. Put $L_\star = \max_i L_i$ and $U_\star = \max_i U_i$. Then

$$\mathcal{F}_{\text{in}} = \{\pi_i : U_\star - L_i \le \varepsilon\}, \qquad \mathcal{F}_{\text{out}} = \{\pi_i : L_\star - U_i \le \varepsilon\} \tag{3}$$

bracket the pipeline-relative near-optimal class on the simultaneous coverage event. We compute diagnostics for both classes and treat differing conclusions as class-uncertain. The outer class accounts for return uncertainty among the enumerated policies; it does not account for missing candidates. Appendix C gives the proof and the corresponding construction when a global value certificate is available.

4 Exact Queries: Separation and Sparse Signatures

Definition 3 (Separated packing and sparse exact signatures)**.** *A family $\Pi = \{\pi_1, \dots, \pi_M\} \subseteq \mathfrak{R}_\varepsilon(\mathcal{M}; \mathcal{F})$ is $2/H$-separated if $\mathrm{TV}(d^{\pi_i}, d^{\pi_j}) > 2/H$ for $i \neq j$. It has b-sparse exact local signatures if, at every state s, some distribution $y_\perp(s)$ satisfies*

$$\left|\{i : \pi_i(\cdot \mid s) \neq y_\perp(s)\}\right| \le b.$$

Separation makes occupancy auditing at radius $1/H$ imply identification of a unique index. Sparsity limits the number of candidates excluded by a background response. Such a response can still convey information; it need not reveal a positive signature to eliminate candidates.

Lemma 4 (Separated packing implies deployment-class capacity)**.** *A $2/H$-separated packing of size M implies $H^{\mathcal{F}}_{\text{opt}}(\varepsilon) \ge \log_2 M$.*

Theorem 5 (Sparse-signature lower bound)**.** *Let I be uniform on $[M]$ and deploy π_I from a separated family with b-sparse exact signatures, $b \ge 1$. Every exact local-query auditor using at most n queries satisfies*

$$\Pr(\widehat{I} = I) \le \min\left\{1, \frac{nb+1}{M}\right\}. \tag{4}$$

Hence success at least $1-\delta$ requires $n \ge ((1-\delta)M - 1)/b$. The same query lower bound holds for randomized minimax occupancy auditing.

To see the counting argument, fix the auditor's randomness and follow its all-background transcript. At most nb indices can first deviate from that transcript. All remaining indices have the same transcript and permit at most one successful final guess. Averaging over the internal randomness preserves the bound. Appendix A supplies the full proof.

Corollary 6 (Capacity form of the search bound)**.** *If the packing in Theorem 5 satisfies $M \ge c\, 2^{H^{\mathcal{F}}_{\text{opt}}(\varepsilon)}$ for a constant $c > 0$ and $b = O(1)$, constant-error occupancy auditing requires $\Omega(2^{H^{\mathcal{F}}_{\text{opt}}(\varepsilon)})$ queries.*

The packing lemma alone gives the opposite inequality, $M \le 2^{H^{\mathcal{F}}_{\text{opt}}(\varepsilon)}$, and is insufficient for this substitution. The weaker condition $M = 2^{\Theta(H^{\mathcal{F}}_{\text{opt}}(\varepsilon))}$ gives only an exponential with a possibly smaller constant in its exponent. Capacity units therefore summarize a realized search family; they do not amplify the search lower bound.

4.1 A realizable hard family and an easy high-capacity control

Theorem 7 (Hidden-branch MDP)**.** *For every $M \ge 2$, a finite discounted MDP with $H = 16M$ admits a class $\mathcal{F}_M$ of M optimal deterministic policies with pairwise occupancy TV greater than $2/H$, 1-sparse exact signatures, and $H^{\mathcal{F}_M}_{\text{opt}}(0) = \log_2 M$. Its optimal Bayes identification success after n exact queries is $\min\{1, (n+1)/M\}$.*

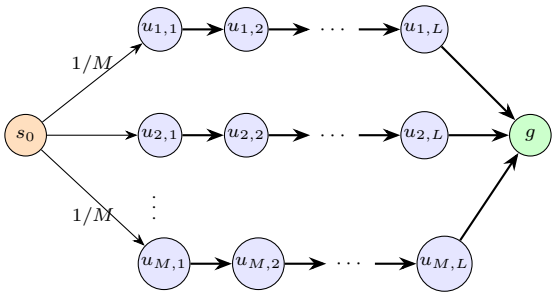


Figure 1: Hidden-branch MDP (Theorem 7). All transitions are essentially deterministic except at s_0, which routes uniformly to one of M branches. Rewards are identically one, so every policy in the deployment class $\mathcal{F}_M = \{\pi_1, \ldots, \pi_M\}$ is optimal. Policy π_i plays the distinguished action only on branch i, producing 1-sparse exact local signatures: querying any branch state reveals which π_i is deployed only when the queried branch happens to coincide with i.

A root routes uniformly to one of M length-H branches and then an absorbing state. Both actions have identical transitions and reward one. Policy π_i uses action one only on branch i. The discounted mass of each branch is $W = \gamma(1-\gamma^H)/M$, so two policies have occupancy TV $2W > 2/H$. One query per branch finds the distinguished action or eliminates that branch. Figure 1 illustrates the construction; Appendix A gives its full proof. The horizon grows with M, and the separation concerns action labels rather than state-marginal trajectories. The full stationary-policy class has additional occupancies; its capacity is not asserted to equal $\log_2 M$.

For an $N \times N$ shortest-path grid, the $M = \binom{2(N-1)}{N-1}$ optimal routes split equally between Right and Down at the start, giving $b \geq M/2$. Path-following identifies the occupancy in $2N-2$ exact queries. Route count alone is not a covering number. Appendix D.1 proves the stronger geometric statement: with $H = 4(N-1)$, a subset of 2^{N-2} optimal routes is $2/H$-separated. Thus this control really has capacity $\Theta(N)$ while admitting an $O(N)$ audit.

4.2 What a supplied occupancy cover does and does not provide

An occupancy cover and a query strategy solve different problems. Suppose an auditor is given a $1/(2H)$-cover $\mathcal{C} = \{d^{\pi_1}, \ldots, d^{\pi_K}\}$ and a finite set Q that separates its representative policies. Separation means that for each pair of representatives some state in Q has different exact action distributions. The further requirement is transcript compatibility: every deployed policy must have a representative within $1/(2H)$ that agrees with it on every query in Q. Under both conditions, the exact transcript identifies a unique representative and returns an occupancy with the required accuracy in $|Q|$ queries (Theorem 15).

The compatibility assumption matters because two nearby occupancies may disagree on a rarely visited or unreachable query state. An arbitrary occupancy cover therefore does not automatically define an exact-transcript lookup table. Nor does receiving a valid cover eliminate search cost: the known hidden-branch family is its own compatible cover, yet identifying its deployed member still takes order K queries. In that construction, one state from each branch gives $|Q| = K$, so the upper and lower bounds match up to constants. Generating a cover is an additional geometric and computational problem; the discounted flow constraints describe feasibility but do not make metric covering efficient. Appendix B retains the complete definitions, proof, and elementary K-bit storage lower bound for arbitrary subsets of a K-member packing.

5 What Changes Under Episode Access?

Root-reset episodes in the hidden-branch MDP independently sample a uniform branch. Observing its state and action reveals whether it is distinguished. If D_n is the number of distinct branches in n episodes, the optimal conditional success is $\min\{1, (D_n+1)/M\}$, giving

$$S_n = 1 - (1-1/M)^n + \frac{\Pr(D_n < M)}{M}. \tag{5}$$

Table 1: Hidden-branch identification at at least 90% Bayes success. State-query budgets use Equation 4; episode budgets use the saturated law in Equation 5. Both scale linearly in M at fixed error.

M	Arbitrary-state exact queries	Root-reset episodes
64	57	137
128	115	284
256	230	579

The last term accounts for the final guess only when an unseen branch remains. Fixed-error identification therefore costs $\Theta(M)$ episodes. Collecting every branch is a different task with coupon-collection cost $\Theta(M \log M)$. Arbitrary-state queries and episodes are not directly ordered in their respective cost units: an episode can provide many observations but cannot choose its visited branch. Table 1 gives exact finite-budget comparisons; Appendix D.2 proves the law.

The derivation exposes the information supplied by an episode. Conditional on a visited set of size $d < M$, each of its d possible distinguished indices can be identified, while the remaining $M - d$ indices have the same background transcript. A final guess contributes one further successful index under the uniform prior. When $d = M$, there is no unseen index to guess, which explains the saturation in Equation 5. Because a branch is missed in every episode with probability $(1 - 1/M)^n$, the expected number of observed branches is $M[1 - (1 - 1/M)^n]$. These two facts determine the finite-budget law without fitting an asymptotic model. For error target δ, $(1-1/M)^n \le \delta$ is sufficient, while $(1 - 1/M)^n \le \delta + 1/M$ is necessary. At fixed error their episode scales coincide.

A partial coverage result applies beyond the hidden-branch construction, provided exact agreement holds throughout the covered region. It is useful to state its conditions explicitly, because average agreement on projected cells is insufficient.

Lemma 8 (Coverage-capped identification)**.** *Let $I \sim \mathrm{Unif}([M])$ select a candidate policy in a common MDP. Episodes start independently from the same μ_0 and stop at an independent geometric time T with $\Pr(T = t) = (1 - \gamma)\gamma^t$. Let $Q \subseteq \mathcal{S}$ be measurable and suppose $d_S^{\pi_i}(Q^c) \le \tau_{\mathrm{cov}}$ for every candidate. Assume the exact action responses at every state in Q have b-sparse signatures. Grant the auditor each visited state and its exact action distribution, with ordinary actions and transitions also observed, and cap the audit at n episodes and q in-region revelations. Then*

$$\Pr(\widehat{I} = I) \le \min\left\{1, \frac{qb+1}{M} + nH\tau_{\mathrm{cov}}\right\}. \tag{6}$$

If the revelation count N_Q is not capped, the valid bound for any deterministic q adds $\Pr(N_Q > q)$ to the right-hand side before clipping at one.

Proof. For candidate i, a geometrically terminated episode has expected number of visits outside Q equal to

$$\sum_{t \ge 0} \Pr(T \ge t) \overset{\pi_i}{\Pr}(s_t \notin Q) = \sum_{t \ge 0} \gamma^t \overset{\pi_i}{\Pr}(s_t \notin Q) = H d_S^{\pi_i}(Q^c) \le H\tau_{\mathrm{cov}}.$$

Across at most n episodes, the probability of any outside visit is at most $nH\tau_{\mathrm{cov}}$ by the union bound and the expected-count bound. Stopping earlier cannot increase this probability.

Compare the actual audit with a censored audit that stops before its first response outside Q. Given the exact in-region responses, its action draws, common-kernel transitions, root resets, and termination can be simulated using randomness independent of I. Fixing this randomness leaves an adaptive exact-query algorithm on Q with at most q responses. The counting argument of Theorem 5 bounds its average success by $\min\{1, (qb + 1)/M\}$. Coupling the two audits makes their outputs identical unless an outside state is visited. Their success probabilities differ by at most that exit probability, giving Equation 6. Censoring additionally at response $q + 1$ contributes at most $\Pr(N_Q > q)$ if no fixed cap was imposed. □

The comparison is unconditional: conditioning on no exit could change the prior on the deployed index. A projected occupancy estimate can bound mass outside covered cells, but exact agreement

must still hold at every concrete state the episode can reveal within those cells. Tolerance agreement or agreement at representatives does not establish this premise. The learned probe statistic in Section 8.1 therefore cannot replace b in Lemma 8 without further structure. Appendix D.3 discusses confidence failure, truncation, and reset assumptions.

6 Noisy Hidden-Trigger Testing

Definition 9 (Noisy hidden-trigger model)**.** *There are M candidate triggers. Under H_0, each sampled trigger returns an independent draw from p_0. Under H_i, only trigger i returns draws from p_1. The alternative is a uniform mixture over $i \in [M]$, and $\beta = \mathrm{KL}(p_0 \| p_1) \in (0, \infty)$.*

Theorem 10 (Noisy hidden-trigger lower bound)**.** *An adaptive test with at most n sample queries and type-I plus mixture type-II error at most $1/3$ must satisfy $n \geq 8M/(9\beta)$. In particular, if $\beta \leq C\rho^2\Delta^2$, then $n = \Omega(M/(\rho^2\Delta^2))$.*

For transcript laws P_0^A, P_i^A and counts N_i, adaptive KL decomposition gives $\mathrm{KL}(P_0^A \| P_i^A) = \beta \mathbb{E}_0 N_i$. Convexity then gives $\mathrm{KL}(P_0^A \| M^{-1} \sum_i P_i^A) \leq n\beta/M$. Pinsker and the testing inequality yield the result; the full proof appears in Appendix A.3. If the alternatives also form a globally near-optimal separated packing with $M \geq c2^{H^{\mathcal{F}}_{\mathrm{opt}}(\varepsilon)}$, the lower bound can be written in capacity units. Detection under the mixture is distinct from identifying an arbitrary deployed occupancy.

The conservative max-scan scale $M \log M/\beta$ is not a universal lower bound. For $p_0 = \delta_0$ and $p_1 = (1-q)\delta_0 + q\delta_1$, $0 < q \leq 1/2$, the signal is $\beta = -\log(1-q) = \Theta(q)$. Cycling through triggers finds a one with constant probability after $\Theta(M/q) = \Theta(M/\beta)$ samples, with zero false positives. Likelihood-ratio structure, error targets, and detection versus localization determine whether a particular scan needs an additional logarithm.

7 Reference Regularization and Its Limits

Reducing the admitted occupancy region can remove a hard packing. Entropy penalties alone do not distinguish reward-tied deterministic policies with zero entropy. The canonical occupancy objective $\widetilde{V}_\lambda(\pi) = V^\pi - \lambda \mathrm{TV}(d^\pi, d^0)$ does: for an optimal reference $\pi_0 \in \mathcal{F}$, every α-near-optimal policy for this objective satisfies $\mathrm{TV}(d^\pi, d^0) \leq \alpha/\lambda$. Thus $\lambda \geq \alpha H$ covers its regularized near-optimal set by one $1/H$ ball. Theorem 18 retains this result and its proof in Appendix G.

A policy-level alternative avoids directly optimizing occupancy TV. Let the reference $\pi_{\mathrm{ref}} \in \mathcal{F}$ satisfy $V^\star - V^{\pi_{\mathrm{ref}}} \leq \eta$, and define

$$K(\pi, \pi_{\mathrm{ref}}) = \mathbb{E}_{s \sim d_S^\pi} \mathrm{KL}\big(\pi(\cdot \mid s) \| \pi_{\mathrm{ref}}(\cdot \mid s)\big), \qquad J_\lambda(\pi) = V^\pi - \lambda K(\pi, \pi_{\mathrm{ref}}). \tag{7}$$

Theorem 11 (Forward policy KL controls occupancy distance)**.** *For policies in the same MDP with finite $K(\pi, \pi_{\mathrm{ref}})$,*

$$\mathrm{TV}(d^\pi, d^{\pi_{\mathrm{ref}}}) \leq \sqrt{HK(\pi, \pi_{\mathrm{ref}})/2}.$$

If $\lambda > 0$ and $J_\lambda(\pi) \geq \sup_{\pi' \in \mathcal{F}} J_\lambda(\pi') - \alpha$, then

$$K(\pi, \pi_{\mathrm{ref}}) \leq \frac{\alpha + \eta}{\lambda}, \qquad \mathrm{TV}(d^\pi, d^{\pi_{\mathrm{ref}}}) \leq \sqrt{\frac{H(\alpha + \eta)}{2\lambda}}.$$

The regularized near-optimal set has capacity zero at radius $1/H$ whenever $\lambda \geq (\alpha + \eta)H^3/2$.

Proof of Theorem 11. Let T be independent of the trajectory, with $\Pr(T = t) = (1-\gamma)\gamma^t$. The terminal pair (s_T, a_T) has law d^π. Consider the complete stopped history $(T, s_0, a_0, \ldots, s_T, a_T)$ under π and π_{ref}, with the same initial distribution and transitions. The geometric-time law, initial-state law, and transition kernels contribute no KL terms. The policy chain rule therefore gives

$$\begin{aligned} \mathrm{KL}(\mathsf{P}_\pi \| \mathsf{P}_{\mathrm{ref}}) &= \sum_{t \geq 0} \Pr(T \geq t) \mathbb{E}_\pi \mathrm{KL}(\pi(\cdot \mid s_t) \| \pi_{\mathrm{ref}}(\cdot \mid s_t)) \\ &= \sum_{t \geq 0} \gamma^t \mathbb{E}_\pi \mathrm{KL}(\pi(\cdot \mid s_t) \| \pi_{\mathrm{ref}}(\cdot \mid s_t)) = HK(\pi, \pi_{\mathrm{ref}}). \end{aligned}$$

This is valid for finite forward KL; reference action support is required wherever the candidate puts positive discounted mass. Data processing to (s_T, a_T) and Pinsker's inequality give

$$\mathrm{TV}(d^{\pi}, d^{\pi_{\mathrm{ref}}}) \le \sqrt{\mathrm{KL}(d^{\pi} \| d^{\pi_{\mathrm{ref}}})/2} \le \sqrt{HK(\pi, \pi_{\mathrm{ref}})/2}.$$

Because $K(\pi_{\mathrm{ref}}, \pi_{\mathrm{ref}}) = 0$ and $\pi_{\mathrm{ref}} \in \mathcal{F}$, the regularized supremum is at least $V^{\pi_{\mathrm{ref}}}$. Thus regularized α-near-optimality gives

$$\lambda K(\pi, \pi_{\mathrm{ref}}) \le V^{\pi} - V^{\pi_{\mathrm{ref}}} + \alpha \le V^{\star} - V^{\pi_{\mathrm{ref}}} + \alpha \le \alpha + \eta.$$

Substitution proves the radius bound. If $\lambda \ge (\alpha + \eta)H^3/2$, the radius is at most $1/H$, and the one ball centered at $d^{\pi_{\mathrm{ref}}}$ covers the regularized near-optimal set. When that set is nonempty its logarithmic cover number is zero. □

The penalty is forward KL weighted by the candidate's occupancy, not reverse KL or reference-weighted KL. Appendix G.1 discusses edge cases and estimation limits.

The reference must support the candidate's actions on visited states. It can be an independently audited policy, a shielded controller, or a checkpoint assessed with stronger tests, but proximity does not make an unaudited reference safe. At $\gamma = 0.99$, complete collapse requires the worst-case sufficient threshold $5 \times 10^5(\alpha+\eta)$. This bound does not explain the empirical reductions at $\lambda \le 0.10$, and the theorem does not guarantee that optimization reaches a specified α.

8 Controlled and Learned-Family Diagnostics

The analytic controls test the conditional structure directly. Tables 5 and 7 instantiate the exact sparse-signature law, while Table 6 provides the high-capacity easy-audit control. Tables 8, 10, and 11 translate the noisy-trigger theorem into declared Bernoulli, continuous-action, and discrete-action signal models. The learned AntMaze experiments in Section 8.2 separately evaluate projected occupancy multiplicity, finite-probe agreement, access restrictions, and regularization. We keep these roles separate: the calibration tables map signal strength to sample scales, whereas the learned-family experiments supply the empirical occupancy and access evidence.

8.1 Separating occupancy evidence from probe evidence

For a fixed finite projection ϕ of state-action pairs, write $p_i = \phi_{\#} d^{\pi_i}$. Assume a simultaneous error event $\mathrm{TV}(\widehat{p}_i, p_i) \le \eta_d$ for every enumerated candidate. An edge between i and j is certified when

$$\mathrm{TV}(\widehat{p}_i, \widehat{p}_j) > 2r + 2\eta_d. \tag{8}$$

Every clique is then a true occupancy packing at separation $2r$, because TV contracts under projection. Its size is a lower bound for the enumerated eligible family, not an estimate of the full cover number. The claim is conditional on valid simultaneous error control, including rollout dependence and truncation.

On an independent finite state-probe set Q, the distinct diagnostic is

$$\widehat{b}^{\mathcal{Y}}_{\tau,Q} = \max_{s \in Q} \min_{y \in \mathcal{Y}_s} \big|\{i : \mathrm{TV}(\pi_i(\cdot \mid s), y) > \tau\}\big|. \tag{9}$$

Restricting the background family $\mathcal{Y}_s$ can only increase this count. If $c_\tau(s)$ candidates at state s are pairwise more than 2τ apart in action TV, then

$$\max_{s \in Q}(c_\tau(s) - 1) \le \widehat{b}^{\mathrm{unrestricted}}_{\tau,Q} \le \widehat{b}^{\mathcal{Y}}_{\tau,Q}. \tag{10}$$

The lower bound follows because a radius-τ ball contains at most one member of that separated subset. Both bounds concern the same finite probe set and tolerance. A finite probe set can miss informative states, and arbitrarily small differences can encode identity under an exact oracle. No uniform exact-sparsity certificate follows from a small value of Equation 9. Appendix E gives the proofs and the required diagnostic records.

For later finite-family summaries we use

$$R_Q = \frac{M_{\mathrm{in}}}{\widehat{b}_{\tau,Q,\mathrm{in}}}$$

Table 2: AntMaze sensitivity analysis on the declared $K = 50$ candidate pool. The packing sweep varies r while holding the return class, projection ϕ, simultaneous projected-occupancy error construction, and candidate pool fixed. The probe sweep varies τ on the same fixed probe set Q. At the default $r = 1/H$ and $\tau = 0.10$, the inner projected packing and probe statistic are $M_{\text{in}} = 20$ and $\widehat{b}_{\tau,Q,\text{in}} = 3$. The rows are separate one-dimensional sensitivity sweeps rather than a complete crossed (r, τ) surface.

Quantity	Setting 1	Setting 2	Setting 3
Packing radius r	$0.5/H$	$1/H$	$2/H$
Inner projected packing $M_{\text{in}}(r)$	20	20	13
Probe tolerance τ	0.05	0.10	0.20
Probe count $\widehat{b}_{\tau,Q}$	4	3	2

only as a dimensionless diagnostic comparing a certified projected packing with probe-local tolerance disagreement. It is not a query-complexity lower bound: $\widehat{b}_{\tau,Q,\text{in}}$ is defined only on the declared finite probe set and at tolerance τ, and it need not equal the exact uniform signature parameter b in Theorem 5.

8.2 Learned families, sensitivity, and access

The learned-family study uses SAC on D4RL `antmaze-medium-play-v2` (Fu et al., 2020) and DQN on MinAtar Breakout (Young & Tian, 2019). AntMaze uses nested pools $K \in \{15, 30, 50\}$, $\varepsilon = 0.05$, $\gamma = 0.99$ ($H = 100$), 200 held-out return rollouts per checkpoint, and family-wise $\delta_V = 0.05$, yielding inner/outer class sizes $(10, 13)$, $(20, 25)$, and $(31, 39)$. Projected occupancy uses 0.5×0.5 position cells with 16 action prototypes fitted on a disjoint calibration split, simultaneous radius $\eta_d = 0.04$, and family-wise $\delta_d = 0.05$.

The independent probe set contains 512 held-out rollout and bottleneck states. At $r = 1/H$ and $\tau = 0.10$, the nested pools have inner projected-packing/probe-count pairs $(7, 2)$, $(14, 3)$, and $(20, 3)$, giving $R_Q = M_{\text{in}}/\widehat{b}_{\tau,Q,\text{in}} = 3.50, 4.67, 6.67$. This dimensionless ratio compares certified projected multiplicity with finite-probe tolerance disagreement; it is not a query lower bound and $\widehat{b}_{\tau,Q,\text{in}}$ is not the exact uniform signature parameter b in Theorem 5.

The SAC policies use five training seeds and 10^6 environment steps per setting. MinAtar uses five DQN seeds and 5×10^6 frames. All learned-family conclusions are pipeline-relative. In particular, even the 50-candidate AntMaze pool has covering number at most 50 and hence capacity at most $\log_2 50 \approx 5.64$ bits; these experiments test the finite-family geometry underlying the theory rather than establish asymptotic exponential auditing hardness.

The AntMaze sensitivity analysis reports inner projected packing sizes $20, 20, 13$ at radii $r \in \{0.5/H, 1/H, 2/H\}$ and probe counts $4, 3, 2$ at tolerances $\tau \in \{0.05, 0.10, 0.20\}$ (Table 2). The packing and probe summaries are separate one-dimensional sweeps rather than a complete crossed surface.

For the access experiment, the deployed index is uniform over the same $K = 50$ candidate pool and the auditor identifies the maximum cumulative action log likelihood, $\arg\max_i \sum_u \log \pi_i(a_u \mid s_u)$. The access conditions differ only in reachability: arbitrary resets use the fixed valid-state bank, observed-state resets use only states reached earlier in the same audit, and trajectory access uses ordinary environment resets and complete on-policy trajectories. At a fixed 20,000-interaction budget, chance identification is 0.02 and success falls from 0.83 ± 0.04 to 0.76 ± 0.05 and 0.61 ± 0.06 as access is restricted. Each value is mean $\pm$ SD over 10 independent batches of 100 deployments; distinct-state counts are medians of unique 0.5×0.5 position cells. Full protocol details are in Appendix F.

At $\lambda = 0$, the outer-class projected packing contains 24 candidates, compared with the 20-member inner packing at $r = 1/H$ in the sensitivity study. Increasing the forward-KL penalty from 0 to 0.10 reduces the outer projected packing from 24 to 3 and the probe statistic from 4 to 1, while normalized return decreases from 0.86 to 0.80. The same projection, packing radius, probe construction, and confidence rule are used throughout, so this is a controlled finite-family return–multiplicity frontier.

Table 3: AntMaze access study on the fixed $K = 50$ candidate family at a 20,000-interaction budget. Success is mean $\pm$ SD over 10 audit batches; cells and episodes are medians across deployments.

Access model	Success	Distinct states	Episodes
Arbitrary valid-state resets	0.83 ± 0.04	412	500
Resets to observed states	0.76 ± 0.05	287	500
Complete on-policy trajectories	0.61 ± 0.06	164	214

Table 4: AntMaze forward-KL return–multiplicity frontier on the fixed $K = 50$ pipeline. All rows use the same projection, confidence construction, $r = 1/H$, probe set, and $\tau = 0.10$; return is mean over five seeds.

λ	Normalized return	M_{out}	$\log_2 M_{\mathrm{out}}$	$\widehat{b}_{0.10,Q,\mathrm{out}}$
0	0.86	24	4.58	4
0.01	0.85	17	4.09	3
0.05	0.83	8	3.00	2
0.10	0.80	3	1.58	1

A decrease in a certified packing lower bound does not prove a corresponding decrease in the full covering number, and the experiment does not establish the worst-case complete-collapse condition of Theorem 11.

Shared training data may encourage agreement on frequently visited states, while reward-equivalent route choices concentrate differences in a few regions. Similar function approximators may also extrapolate similarly outside the training distribution. These remain mechanism hypotheses. A diversity-regularized learned control and independent expansion of the probe set would test them: persistent small probe counts after deliberate behavioral diversification would weaken the shared-behavior explanation. The analytic grid already supplies a non-sparse control; the learned interventions are not reported as completed experiments.

9 Discussion and Conclusion

Our results separate behavioral multiplicity from the information available to distinguish policies. A large near-optimal family or large occupancy covering number does not by itself imply a difficult audit. Under b-sparse exact local signatures, however, an occupancy-separated family of size M requires $\Omega(M/b)$ exact queries. The hidden-branch construction realizes this regime, while the shortest-path control shows that high occupancy capacity can still be easy to audit when observations partition candidates informatively. Thus the capacity form of the lower bound should be interpreted as a geometric description of a realized hard family, not as a universal consequence of large policy multiplicity.

Access restrictions provide a second source of difficulty. Arbitrary-state queries can target informative regions directly, whereas trajectory-based auditing is constrained by reachability. The AntMaze study reflects this distinction: identification falls from 0.83 with arbitrary valid-state resets to 0.76 with observed-state resets and 0.61 under complete trajectories at the same interaction budget. These results remain finite-family diagnostics rather than empirical verification of the sparse-signature theorem. In particular, the default $20/3 = 6.67$ packing-to-probe ratio compares certified projected multiplicity with finite-probe tolerance disagreement, but it is not itself a query lower bound.

The regularization results provide a complementary mitigation view. Increasing the AntMaze forward-KL penalty from 0 to 0.10 reduces the outer projected packing from 24 to 3 and the probe statistic from 4 to 1, while normalized return decreases from 0.86 to 0.80. This demonstrates a finite-family return–multiplicity tradeoff, not the theorem's worst-case capacity-collapse condition. Overall, meaningful auditing claims should report the deployment class, occupancy resolution, access model, and local response structure together. Candidate count or return similarity alone is insufficient to characterize auditing difficulty.

AI USE STATEMENT

We used OpenAI ChatGPT as a writing and editing aid. It helped improve the clarity of the exposition, cross-check notation and the internal consistency of stated results, recover bibliographic metadata, and diagnose LaTeX compilation and layout issues. The authors developed the research questions, definitions, theorems, proofs, constructions, and experimental design. The authors checked every proof and every reported number. All analytic tables are regenerated from the listings in Appendix I, not taken from model output. No language model was used to generate, modify, or select experimental measurements, and none is a component of the auditing methods studied. The authors take full responsibility for the content of this paper.

ETHICS STATEMENT

This work asks when near-optimal reinforcement-learning policies can be audited. The motivating concern is that policies with equivalent reward may hide divergent or trigger-conditioned behavior. The work involves no human subjects and no personal or sensitive data. Experiments use simulated MDPs and the public D4RL AntMaze and MinAtar benchmarks.

The main dual-use consideration is that describing hard-to-audit families also describes how behavior could be concealed. We consider this risk limited, for three reasons. The hidden-branch and trigger-room constructions are abstract instances of hidden-index search. Sparse trigger-conditioned behavior is already documented in the RL backdoor literature. We give no procedure for inserting triggers into learned policies.

The intended use is defensive. The results tell auditors which assumptions a claim of auditability must state: the deployment class, occupancy resolution, access model, and local response structure. They also caution against treating return similarity or candidate counts as evidence of safety. A second risk is overinterpretation, which we address in the main text. The learned-family diagnostics are finite and pipeline-relative, so they do not certify the absence of hidden behavior, and a small probe statistic does not establish uniform agreement. Likewise, the reference-regularization results reduce multiplicity only relative to a reference policy whose trustworthiness must be established independently.

REPRODUCIBILITY STATEMENT

Every theoretical result states its assumptions where it appears. Complete proofs are in Appendix A, which covers the sparse-signature bound, the hidden-branch construction, and the noisy-trigger bound. The remaining proofs are in Appendix D.1 (grid capacity), Appendix D.2 (root-reset law), and Appendix G (occupancy regularizer). The proofs of Lemma 8 and Theorem 11 are in the main text. Appendix I contains self-contained Python listings, using only the standard library, that deterministically regenerate every analytic table: Tables 1, 5, 6, 7, 8, 9, 10, and 11. The learned-family experiments are specified in Section 8.2 and Appendices E and F.

REFERENCES

Chace Ashcraft and Kiran Karra. Poisoning deep reinforcement learning agents with in-distribution triggers, 2021. URL https://arxiv.org/abs/2106.07798.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J. Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback, 2023. URL https://arxiv.org/abs/2307.15217.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: Datasets for deep data-driven reinforcement learning, 2020. URL https://arxiv.org/abs/2004.07219.

Nathan Fulton and André Platzer. Safe reinforcement learning via formal methods: Toward safe control through proof and learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Mohammadhosein Hasanbeig, Alessandro Abate, and Daniel Kroening. Cautious reinforcement learning with logical constraints. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 483–491, 2020.

Hsiang Hsu and Flavio Calmon. Rashomon capacity: A metric for predictive multiplicity in classification. *Advances in Neural Information Processing Systems*, 35:28988–29000, 2022.

Panagiota Kiourti, Kacper Wardega, Susmit Jha, and Wenchao Li. Trojdrl: evaluation of backdoor attacks on deep reinforcement learning. In *2020 57th ACM/IEEE Design Automation Conference (DAC)*, pp. 1–6. IEEE, 2020.

Saurabh Kumar, Aviral Kumar, Sergey Levine, and Chelsea Finn. One solution is not all you need: Few-shot extrapolation via structured maxent rl. *Advances in Neural Information Processing Systems*, 33:8198–8210, 2020.

Lucien LeCam. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pp. 38–53, 1973.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *International conference on machine learning*, pp. 6765–6774. PMLR, 2020.

Sebastian Müller, Vanessa Toborek, Katharina Beckh, Matthias Jakobs, Christian Bauckhage, and Pascal Welke. An empirical evaluation of the rashomon effect in explainable machine learning. In *Joint european conference on machine learning and knowledge discovery in databases*, pp. 462–478. Springer, 2023.

Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models, 2022. URL https://arxiv.org/abs/2201.03544.

Amin Rakhsha, Goran Radanovic, Rati Devidze, Xiaojin Zhu, and Adish Singla. Policy teaching via environment poisoning: Training-time adversarial attacks against reinforcement learning. In *International Conference on Machine Learning*, pp. 7974–7984. PMLR, 2020.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys*, 16:1–85, 2022.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1827–1858, 2022.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward gaming. *Advances in Neural Information Processing Systems*, 35:9460–9471, 2022.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, New York, NY, 2009. doi: 10.1007/b13794.

Lun Wang, Zaynah Javed, Xian Wu, Wenbo Guo, Xinyu Xing, and Dawn Song. Backdoorl: Backdoor attack against competitive reinforcement learning. *arXiv preprint arXiv:2105.00579*, 2021.

Andrew Chi-Chih Yao. Should tables be sorted? *Journal of the ACM (JACM)*, 28(3):615–628, 1981.

Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pp. 222–227. IEEE, 1977.

Kenny Young and Tian Tian. MinAtar: An Atari-inspired testbed for thorough and reproducible reinforcement learning experiments, 2019. URL https://arxiv.org/abs/1903.03176.

Bin Yu. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pp. 423–435. Springer, 1997.

A Additional Proof Details

A.1 Packing, exact signatures, and randomized auditors

No closed TV ball of radius $1/H$ contains two points at distance greater than $2/H$, by the triangle inequality. Covering a separated packing therefore requires at least one ball per member, proving Lemma 4. The same argument works at any radius $r > 0$ with separation strictly greater than $2r$.

Proof of Theorem 5. Fix a deterministic auditor and its all-background query path. Let B be the union of the indices that have a non-background response to any state on that path. There are at most nb such indices. If the deployed index is outside B, induction on the queries shows that the auditor follows the all-background path, so its final output can identify at most one such index. At most $|B|$ indices can be identified on all other transcripts, including arbitrarily adaptive behavior after the first non-background response. Hence the number of successful indices is at most $nb + 1$, and the Bayes success is at most $\min\{1, (nb+1)/M\}$. Integrating this bound over the auditor's random seed proves it for randomized auditors. A worst-case success guarantee implies the same guarantee under the uniform prior, giving the minimax lower bound. Finally, occupancy accuracy within $1/H$ identifies at most one packing member, so the identification bound applies to occupancy auditing. □

This is the standard Yao reduction: a randomized algorithm that beat the bound on every index would have a random seed beating it on average under the uniform prior. When $M = 1$ identification is free. A separated family with $M > 1$ cannot have $b = 0$, since identical action distributions at every state induce identical occupancies in the common MDP.

A.2 Hidden-branch construction and constants

Proof of Theorem 7. Let $H = 16M$, $\gamma = 1 - 1/H$, and $L = \lceil H \rceil$. The state space consists of a root s_0, an absorbing state g, and branch states $u_{j,\ell}$ for $j \in [M]$ and $\ell \in [L]$; the action set is $\mathcal{A} = \{0, 1\}$ and the initial distribution is concentrated on s_0. From s_0, both actions transition uniformly to $u_{j,1}$ over $j \in [M]$. For $\ell < L$, both actions at $u_{j,\ell}$ transition to $u_{j,\ell+1}$; from $u_{j,L}$, both actions transition to g, and g is absorbing. Rewards are identically equal to one, so every policy has value one and every policy in $\mathcal{F}_M$ is optimal. For each $i \in [M]$, π_i plays action 1 at every state $u_{i,\ell}$ on branch i and action 0 at all other branch states; actions at s_0 and g are fixed identically across policies. Querying a state on branch j returns action 1 if and only if $i = j$, so the signatures are 1-sparse.

Setting $W = \frac{1}{M}\sum_{\ell=1}^{L}(1-\gamma)\gamma^{\ell} = \frac{\gamma(1-\gamma^L)}{M}$ and using $\gamma \geq 1/2$, $1 - \gamma^L \geq 1 - e^{-1}$ for $H \geq 2$, we have $W \geq (1-e^{-1})/(2M)$. For $i \neq j$, π_i and π_j differ on branches i and j, giving $\mathrm{TV}(d^{\pi_i}, d^{\pi_j}) = 2W \geq (1 - e^{-1})/M > 2/H = 1/(8M)$. Since $\mathcal{F}_M$ contains exactly these M pairwise $2/H$-separated occupancies, the cover number at radius $1/H$ is M, so $H_{\mathrm{opt}}^{\mathcal{F}_M}(0) = \log_2 M$, and the query lower bound follows from Theorem 5 with $b = 1$. □

The matching upper bound queries distinct branches and, after $M-1$ background responses, identifies the remaining index without querying it. For $n < M$, the queried indices and a single fallback index give success $(n + 1)/M$. The choice $H = 16M$ is sufficient, not necessary: $H = cM$ with a sufficiently large constant and branch length $L = \Theta(H)$ preserves branch mass $\Theta(1/M)$ and audit radius $\Theta(1/M)$. The state count in the stated construction is $2 + ML = O(M^2)$. The result does not hold a fixed horizon or a fixed state count while sending M to infinity.

A.3 Noisy hidden-trigger proof and error constants

Proof. Fix any adaptive algorithm using at most n samples, and let P_0^A, P_i^A be the transcript laws under H_0 and H_i, with N_i the number of samples drawn from trigger i. The chain rule for KL divergence under adaptive sampling gives $\mathrm{KL}(P_0^A \| P_i^A) = \mathbb{E}_0[N_i]\beta$. Let $P_{\mathrm{mix}}^A = M^{-1}\sum_i P_i^A$ be the mixture transcript law. Convexity of $Q \mapsto \mathrm{KL}(P_0^A \| Q)$ in its second argument yields

$$\mathrm{KL}(P_0^A \| P_{\mathrm{mix}}^A) \leq \frac{1}{M}\sum_{i=1}^{M}\mathrm{KL}(P_0^A \| P_i^A) = \frac{\beta}{M}\sum_{i=1}^{M}\mathbb{E}_0[N_i] \leq \frac{n\beta}{M},$$

and Pinsker's inequality gives $\mathrm{TV}(P_0^A, P_{\mathrm{mix}}^A) \le \sqrt{n\beta/(2M)}$. The sum of type-I and type-II errors of any test is at least $1 - \mathrm{TV}(P_0^A, P_{\mathrm{mix}}^A)$, so if $n \le cM/\beta$ for a sufficiently small universal constant c, this lower bound exceeds $1/3$. More explicitly, an error sum at most $1/3$ requires $\mathrm{TV}(P_0^A, P_{\mathrm{mix}}^A) \ge 2/3$, hence $n \ge 8M/(9\beta)$. A capacity substitution also requires $M \ge c2^{H_{\mathrm{opt}}^{\mathcal{F}}(\varepsilon)}$ as in Corollary 6; separation alone is insufficient. □

For randomized tests, internal randomness is included in the transcript and has the same distribution under every hypothesis. A bounded stopping time can be padded with uninformative observations to give the fixed-budget argument. If $\beta = 0$, the hypotheses have identical sample laws and detection is impossible; if $\beta = \infty$, this finite-KL bound is uninformative.

For the revealing-event example in Section 6, take $k = \lceil \log(1/\delta)/q \rceil$ complete cycles. Under any alternative the probability of missing the event is $(1-q)^k \le \delta$; under the null no one is ever observed. A one also identifies the anomalous trigger. For $0 < q \le 1/2$, $q \le -\log(1-q) \le 2q$, so constant error is achieved with $O(M/\beta)$ samples. This establishes that a universal extra $\log M$ cannot be inferred from the single scalar β.

B Static Recognition and Oracle-Cover Verification

The lower bounds above describe what an auditor cannot do under sparse signatures or weak per-sample signals. To complete the picture we record two complementary statements: a static information bound that explains why arbitrary target subsets of a finite packing require one bit per packing member, and an oracle-cover verification result that separates the verification half of the auditing problem from cover generation. The supplied query set determines the verification cost. In the hidden-branch family this cost is linear in the number of representatives, matching the search lower bound up to constants.

Theorem 12 (Static target-recognition information lower bound)**.** *Let $\Pi = \{\pi_1, \dots, \pi_K\}$ be a near-optimal occupancy packing. A deterministic exact data structure that preprocesses an arbitrary target subset $T \subseteq \Pi$ and answers membership queries "is $\pi_i \in T$?" for every $i \in [K]$ must use at least K bits of memory in the worst case.*

Proof. There are 2^K possible target subsets of Π. A deterministic exact data structure using fewer than K bits has fewer than 2^K memory states, so two distinct target subsets must share a memory representation; they differ on at least one query index i, forcing an incorrect answer for that query. □

The argument is elementary and is included only to register that exact descriptions of arbitrary subsets of a K-member packing have unavoidable linear cost; the bound is in the cell-probe spirit but does not pursue refined time–space tradeoffs.

Definition 13 (Separating query set)**.** *A finite set of states $Q \subseteq \mathcal{S}$ separates a cover $\mathcal{C} = \{d^{\pi_1}, \dots, d^{\pi_K}\}$ if, for every $i \ne j$, there exists $s \in Q$ such that $\pi_i(\cdot \mid s) \ne \pi_j(\cdot \mid s)$.*

Definition 14 (Transcript-compatible cover)**.** *A cover $\mathcal{C} = \{d^{\pi_1}, \dots, d^{\pi_K}\}$ of $\mathcal{D}_\varepsilon(\mathcal{M}; \mathcal{F})$ is Q-transcript-compatible if, for every deployed policy $\pi \in \mathfrak{R}_\varepsilon(\mathcal{M}; \mathcal{F})$, there exists an index $j \in [K]$ such that*

$$\mathrm{TV}(d^{\pi}, d^{\pi_j}) \le \frac{1}{2H} \qquad \text{and} \qquad \pi(\cdot \mid s) = \pi_j(\cdot \mid s) \quad \text{for all } s \in Q.$$

Theorem 15 (Oracle-cover verification upper bound)**.** *Suppose an auditor is given a $1/(2H)$-cover $\mathcal{C} = \{d^{\pi_1}, \dots, d^{\pi_K}\}$ of $\mathcal{D}_\varepsilon(\mathcal{M}; \mathcal{F})$ and a finite query set $Q \subseteq \mathcal{S}$ such that:*

1. *Q separates the representative policies, meaning that for every $i \ne j$ there exists $s \in Q$ with $\pi_i(\cdot \mid s) \ne \pi_j(\cdot \mid s)$; and*
2. *$\mathcal{C}$ is Q-transcript-compatible in the sense of Definition 14.*

Then, under the exact local-query oracle, querying every state in Q identifies a unique cover representative π_j whose transcript agrees with the deployed policy on Q, using $|Q|$ queries. The returned occupancy d^{π_j} satisfies

$$\mathrm{TV}(d^{\pi_j}, d^{\pi}) \le \frac{1}{2H} \le \frac{1}{H}.$$

In the hidden-branch construction, one can take $|Q| = K$, matching Theorem 5 up to constants.

Proof. Querying all states in Q gives the exact transcript $\{\pi(\cdot \mid s) : s \in Q\}$ of the deployed policy. By Q-transcript-compatibility, there exists at least one representative π_j whose transcript agrees with the deployed policy on every state in Q and whose occupancy lies within $1/(2H)$ of d^π. By the separating property of Q, no two distinct representatives can have the same transcript on Q, so this representative is unique. Returning d^{π_j} therefore achieves

$$\mathrm{TV}(d^{\pi_j}, d^\pi) \le \frac{1}{2H} \le \frac{1}{H}.$$

The query count is exactly $|Q|$.

For the hidden-branch construction, take one query state from each branch, for example

$$Q = \{u_{j,1} : j \in [K]\}.$$

The transcript on these states identifies which branch carries the distinguished action, so Q separates the K representatives and has size K. □

Theorem 15 is intentionally a verification statement and does not claim that valid covers can be generated efficiently for arbitrary MDPs. Cover generation is a separate geometric and computational problem; in tabular MDPs the feasible occupancy region is described by the flow constraints of Section 3, but exact covering may itself be exponential in the capacity. Theorem 15 separates geometric cover construction from transcript verification; it does not make either step generally cheap. In particular, even a fully known hidden-branch candidate family needs $\Omega(K)$ exact queries for constant-error identification. Occupancy proximity alone does not supply transcript compatibility.

C Deployment-Class Construction and Return Uncertainty

A pipeline manifest specifies the finite policy family before the audit: environment version, reward and observation wrappers, algorithm and hyperparameters, training budget, seed list, checkpoint-selection rule, checkpoint hashes, preprocessing, and the stochastic or deterministic action rule used at deployment. Distinct checkpoint files that induce the same occupancy are one occupancy class. Conversely, common architecture or similar return does not make distinct occupancies equivalent. All class-dependent conclusions apply only to the declared candidates. When the pipeline is incompletely known, an outer return class must not be interpreted as a completeness certificate for its outputs.

Let $\mathcal{E}_V = \{L_i \le V^{\pi_i} \le U_i \text{ for every } i\}$ be a simultaneous confidence event with probability at least $1 - \delta_V$. Intervals may be constructed by a valid bounded-return method with a family-wise allocation; the theorem requires simultaneous coverage and does not endorse a particular method for dependent rollout data. Define $V^\star_{\mathcal{P}} = \max_i V^{\pi_i}$ and $\mathcal{R}^{\mathcal{P}}_\varepsilon = \{\pi_i : V^\star_{\mathcal{P}} - V^{\pi_i} \le \varepsilon\}$.

Proposition 16 (Pipeline-relative return bracketing)**.** *On $\mathcal{E}_V$, Equation 3 satisfies $\mathcal{F}_{\mathrm{in}} \subseteq \mathcal{R}^{\mathcal{P}}_\varepsilon \subseteq \mathcal{F}_{\mathrm{out}}$.*

Proof. On the simultaneous event, $L_\star \le V^\star_{\mathcal{P}} \le U_\star$. If $\pi_i \in \mathcal{F}_{\mathrm{in}}$, then $V^\star_{\mathcal{P}} - V^{\pi_i} \le U_\star - L_i \le \varepsilon$. If $\pi_i \in \mathcal{R}^{\mathcal{P}}_\varepsilon$, then $L_\star - U_i \le V^\star_{\mathcal{P}} - V^{\pi_i} \le \varepsilon$, so it belongs to $\mathcal{F}_{\mathrm{out}}$. □

When a valid upper certificate $U_{\mathrm{glob}} \ge V^\star$ is available, $\{\pi_i : U_{\mathrm{glob}} - L_i \le \varepsilon\}$ is a globally near-optimal inner class. The outer rule based on $L_\star - U_i \le \varepsilon$ contains every globally near-optimal enumerated policy, since $L_\star \le V^\star_{\mathcal{P}} \le V^\star$. These global and pipeline-relative inner classes need not coincide. The learned diagnostics use the latter and retain that qualification even when their projected separation is certified.

D Additional Access and Geometry Results

D.1 A high-capacity shortest-path family

We make the geometric content of the grid control explicit rather than equating the count of routes with a covering number. Let $m = N - 1 \ge 2$, use coordinates $\{0, \dots, m\}^2$, and take $H = 4m$ and

$\gamma = 1 - 1/H$. A route starts at $(0, 0)$ and ends at (m, m); rewards are zero before the goal and one at its absorbing continuation. Every shortest route takes $2m$ actions and has the same optimal discounted return. Complete each route policy with fixed actions off the route and at the goal.

Within the $\binom{2m}{m}$ shortest routes, consider those formed by m successive two-step choices: at diagonal state (k, k), take either Right–Down or Down–Right to reach $(k + 1, k + 1)$. Index these routes by binary strings of length m and retain the even-parity strings. There are 2^{m-1} of them and distinct strings differ in at least two gadgets. Every differing gadget contributes two nonoverlapping state-action atoms, each with discounted mass at least $(1 - \gamma)\gamma^{2m-1}$. Therefore distinct retained routes satisfy

$$\mathrm{TV}(d^{\pi_i}, d^{\pi_j}) \geq 4(1 - \gamma)\gamma^{2m-1} > 2/H.$$

The strict inequality follows from Bernoulli's inequality: $\gamma^{2m-1} \geq 1 - (2m - 1)/(4m) > 1/2$. Thus the full shortest-route deployment family has capacity at least $m - 1$, while covering each route separately gives capacity at most $\log_2 \binom{2m}{m} = O(m)$. This proves $\Theta(N)$ capacity at the stated discounted resolution. Querying the successive visited states still identifies the route in $2m = 2N - 2$ queries.

For the full route family, exactly half of the policies take each of the two initial actions, so every exact common-response choice has at least $M/2$ exceptions at the start. For the even-parity subfamily the same balanced split holds for $m \geq 2$. This is a failure of the sparse-signature hypothesis, not a counterexample to Theorem 5. The finite-horizon enumeration in Table 9 uses a different, explicitly stated occupancy normalization and is not used to identify the discounted covering number.

D.2 Root-reset Bayes success

Proposition 17 (Root-reset identification law)**.** *In the hidden-branch construction, after n independent root-reset episodes observing branch identities and actions, the optimal Bayes success under the uniform prior is Equation 5.*

Proof. The visited branch set B_n is independent of the deployed index I. Conditional on B_n with $|B_n| = d < M$, a distinguished observation identifies each of the d visited indices, while all $M - d$ unvisited indices have the same all-zero transcript. Their contribution to unconditional success is one additional index, so optimal conditional success is $(d + 1)/M$. If $d = M$, success is one, rather than $(M + 1)/M$. Also $\mathbb{E}D_n = M(1 - (1 - 1/M)^n)$, since each branch is missed with probability $(1 - 1/M)^n$. Taking the expectation of the saturated conditional success yields

$$S_n = \frac{\mathbb{E}D_n + \Pr(D_n < M)}{M} = 1 - (1 - 1/M)^n + \frac{\Pr(D_n < M)}{M}.$$

The posterior-optimal rule achieves this value by identifying a positive branch or guessing uniformly among the unseen branches. □

Writing $a_n = (1 - 1/M)^n$, we have $1 - a_n \leq S_n \leq 1 - a_n + 1/M$. For any fixed target error $\delta \in (0, 1)$, these inequalities imply the necessary condition $a_n \leq \delta + 1/M$ and the sufficient condition $a_n \leq \delta$. Both give $\Theta(M)$ episodes as $M \to \infty$. Full branch coverage has a different objective and the coupon-collector scale $\Theta(M \log M)$. The displayed budgets in Table 1 use the exact occupancy-count recursion, not an unsaturated approximation or a fitted curve. Episodes here can be truncated after their first branch observation, which already provides all information about that branch; episode counts are not interaction-step counts.

D.3 Identification with incomplete state coverage

The full statement and proof of Lemma 8 appear in Section 5.

The argument uses an unconditional coupling. Conditioning directly on no exit can change the index prior and is insufficient for the counting step. The common transition model is an information assumption; it need not be computationally cheap for an auditor to simulate. Geometric termination gives exactly the discounted visit measure. A deterministic truncation only decreases the number of visits, but arbitrary resets to other initial distributions need a separate coverage assumption for those distributions.

If Q is a union of projected state cells, a simultaneous bound on mass outside these cells can supply τ_{cov}. It supplies no statement about variation within a cell. The exact-sparsity premise must still hold at every concrete state that can be revealed. A finite tolerance statistic with $\tau > 0$, a cell-average action law, or agreement at a representative state does not establish this premise. If confidence events jointly establish the premises with failure probability δ_{cert}, an unconditional bound adds δ_{cert}. No numerical coverage certificate or complete baseline bracket is reported for the learned families here.

D.4 Other information structures

Suppose every residual candidate set of size greater than one admits an exact query whose response classes each contain at most a $(1 - c)$ fraction of that set, for a fixed $c > 0$. Following the observed response then reduces the residual size geometrically, identifying the index in at most $\lceil \log M / [-\log(1 - c)] \rceil$ queries. Such recursively balanced splitting is incompatible with an order-M worst-case search requirement. Sparse signatures are a concrete sufficient form of persistent imbalance, not a characterization of every hard family. Common responses with at most b exceptions give the exact-query theorem; weak likelihood separation gives the noisy lower bound; and small visitation probability can create an additional reachability cost under episode access. These conditions concern different information restrictions and should not be merged into one empirical statistic.

E Projected Occupancy and Local-Response Diagnostics

For continuous-control diagnostics, occupancies are probability measures and state-action sums are interpreted as integrals. The projection, stopping-time, and KL arguments continue to hold for measurable policies and common kernels with finite stated divergences. Fix a measurable finite projection ϕ before evaluation, or establish its confidence bounds conditionally on an independent fitting split. Let $p_i = \phi_{\#} d^{\pi_i}$, and let $\mathcal{E}_d$ be the event that $\mathrm{TV}(\widehat{p}_i, p_i) \le \eta_d$ simultaneously for every candidate. The triangle inequality gives

$$\mathrm{TV}(p_i, p_j) \ge \mathrm{TV}(\widehat{p}_i, \widehat{p}_j) - 2\eta_d.$$

By data processing, $\mathrm{TV}(d^{\pi_i}, d^{\pi_j}) \ge \mathrm{TV}(p_i, p_j)$. Consequently, the single margin in Equation 8 certifies every pair in the selected clique on $\mathcal{E}_d$, even when the clique is chosen after seeing all pairwise distances. One must not subtract the occupancy error twice and then require the same margin again. A clique-finding heuristic yields a valid lower bound if its edges are certified; maximality or maximum-clique optimality is not needed for this conclusion.

Apply this construction separately to $\mathcal{F}_{\mathrm{in}}$ and $\mathcal{F}_{\mathrm{out}}$. On $\mathcal{E}_V \cap \mathcal{E}_d$, an inner-class clique is a certified separated subset of the pipeline-relative near-optimal region. An outer-class clique is separated among possibly eligible candidates; it is not a certified near-optimal packing. Neither class is globally near-optimal without a global value certificate. Projection can hide occupancy distinctions, so failure to find a large clique does not upper-bound full occupancy capacity.

For the baseline bracket, let C_s be any $c_\tau(s)$-member subset of candidate action distributions that is pairwise more than 2τ apart. A distribution y cannot be within τ of two members of C_s, because that would put them within 2τ of each other. Thus at least $c_\tau(s) - 1$ members lie outside every such ball. Minimizing over all y and then maximizing over $s \in Q$ gives the lower bound in Equation 10. Restricting the minimization to $\mathcal{Y}_s$ gives its upper bound. If action laws themselves are estimated, their error must also enter the pairwise separation and tolerance comparisons; the deterministic bracket is stated for the evaluated distributions.

For the learned-family analysis, the candidate manifest records pool membership, training seed, checkpoint selection, and deployed action rule. Return bracketing uses the declared ε, held-out rollout count, and family-wise error level δ_V. Projected occupancy uses the fixed mapping ϕ, a calibration split disjoint from occupancy evaluation, simultaneous error radius η_d, and family-wise error level δ_d. The probe record contains every state in Q, its source (occupancy-weighted or bottleneck), the candidate action distributions, the background optimizer, and the tolerance τ. Packing edges and cliques are computed only after applying the margin in Equation 8. These records make the projected packing interpretation reproducible while preserving the distinction between finite-probe tolerance agreement and uniform exact sparsity.

For an unrestricted expressive policy class over a continuous state space, finitely many probes cannot certify uniform exact agreement: policies can agree at every probed state and differ on an unprobed region. Additional regularity, exhaustive finite-state checking, or a formal within-cell guarantee can change that conclusion. A K-checkpoint pool also has occupancy covering number at most K, since each candidate can be its own center. These facts limit what the finite learned study establishes without invalidating a certified separated subset of the enumerated family.

F Learned-Family Results and Sensitivity

The nested AntMaze/SAC candidate pools contain 15, 30, and 50 policies. At the default $r = 1/H$ and $\tau = 0.10$, the corresponding inner-class projected packing sizes are 7, 14, and 20, while the probe statistics are 2, 3, and 3. Their descriptive ratios $M_{\mathrm{in}}/\widehat{b}_{\tau,Q,\mathrm{in}}$ are therefore 3.50, 4.67, and 6.67.

The access experiment uses the $K = 50$ pool. Each condition is evaluated with 10 independent audit batches of 100 uniformly drawn deployments per batch. The reported 0.83 ± 0.04, 0.76 ± 0.05, and 0.61 ± 0.06 values are mean $\pm$ SD of batch-level identification accuracy; the chance rate is 0.02. Distinct-state counts are medians of unique 0.5×0.5 agent-position cells visited within the fixed interaction budget.

The normalized returns in the forward-KL study are 0.86 ± 0.02, 0.85 ± 0.02, 0.83 ± 0.03, and 0.80 ± 0.03 at $\lambda \in \{0, 0.01, 0.05, 0.10\}$, where $\pm$ denotes SD across the five independently trained seeds. The projection, packing radius, simultaneous confidence construction, probe set, and held-out return procedure are fixed across penalties. The experiment therefore measures a pipeline-relative return–projected-multiplicity frontier; it does not upper-bound the full occupancy covering number or verify the worst-case collapse threshold of Theorem 11.

For the access experiment, the deployed index is drawn uniformly from the $K = 50$ candidate pool. The auditor maintains for every candidate the cumulative action log likelihood

$$\ell_i(t) = \sum_{u=1}^{t} \log \pi_i(a_u \mid s_u)$$

and returns $\arg\max_i \ell_i$ at the interaction budget. The three access conditions differ only in which states may be reached. Arbitrary valid-state reset may initialize an episode at any state in the fixed valid-state bank; observed-state reset may initialize only at a state reached previously in the same audit; and the trajectory condition uses ordinary environment resets and complete on-policy trajectories. Thus the identification rule and candidate family are held fixed while state access is restricted.

At a fixed 20,000 interaction-step budget, the access study uses the same $K = 50$ candidate family under every access model. Chance identification under the uniform deployment prior is therefore $1/K = 0.02$. Each reported success rate is the mean over 10 independent audit batches, each containing 100 independently sampled deployments; the $\pm$ quantity is the standard deviation of the batch-level identification rate. Arbitrary valid-state resets achieve 0.83 ± 0.04, resets to previously observed states achieve 0.76 ± 0.05, and complete on-policy trajectories achieve 0.61 ± 0.06.

Because AntMaze has a continuous state space, "distinct states" means distinct 0.5×0.5 agent-position cells, not exact floating-point observations. We report the median number of distinct visited cells and completed episodes across deployments. The comparison therefore measures the finite-sample effect of restricting reachability while holding the candidate family, auditor, and interaction budget fixed; it is not a generic trajectory-query lower bound.

Two prospective checks distinguish the mechanism hypotheses. First, a learned family trained for deliberate behavioral diversity should create broader local disagreement; continued small counts on adequately expanded probes would weaken the shared-background explanation. Second, adding independent occupancy-weighted and bottleneck probes tests whether earlier small counts arose from incomplete coverage. The first intervention changes the learned family, while the second changes the observed region at a fixed family. Neither is reported as a completed validation experiment.

G REFERENCE-REGULARIZATION PROOFS

Define the regularized near-optimal occupancy set

$$\widetilde{\mathcal{D}}^{\mathcal{F}}_{\alpha,\lambda} = \left\{ d^\pi : \pi \in \mathcal{F},\ \widetilde{V}_\lambda(\pi) \geq \sup_{\pi' \in \mathcal{F}} \widetilde{V}_\lambda(\pi') - \alpha \right\}.$$

The corresponding regularized audited Rashomon capacity is

$$\widetilde{H_{\mathrm{opt}_\lambda}}^{\mathcal{F}}(\alpha) = \log_2 \mathcal{N}\left(\widetilde{\mathcal{D}}^{\mathcal{F}}_{\alpha,\lambda}, \frac{1}{H}, \mathrm{TV} \right).$$

Theorem 18 (Canonical occupancy regularizer)**.** *Assume there exists an audited optimal reference policy $\pi_0 \in \mathcal{F}$ with occupancy $d^0 = d^{\pi_0}$ and value $V^{\pi_0} = V^\star$. Define $\widetilde{V}_\lambda(\pi) = V^\pi - \lambda \mathrm{TV}(d^\pi, d^0)$. Let $\lambda > 0$. If $\pi \in \mathcal{F}$ is α-near-optimal for the regularized objective, $\widetilde{V}_\lambda(\pi) \geq \widetilde{V}_\lambda(\pi_0) - \alpha$, then $\mathrm{TV}(d^\pi, d^0) \leq \alpha/\lambda$. Consequently, if $\lambda \geq \alpha H$, the regularized near-optimal occupancy set over $\mathcal{F}$ lies inside a single $1/H$-ball and has audited Rashomon capacity zero at tolerance α.*

Proof. Since $\widetilde{V}_\lambda(\pi_0) = V^\star$ and $V^\pi \leq V^\star$ for every policy, the regularized near-optimality condition gives $\lambda \mathrm{TV}(d^\pi, d^0) \leq \alpha - (V^\star - V^\pi) \leq \alpha$. The capacity conclusion follows because all regularized near-optimal occupancies are covered by the single ball centered at d^0 with radius $1/H$. □

The mitigation is mathematical, not free in practice: the reference occupancy must itself be trusted, and scalable implementations require reliable empirical surrogates for occupancy distance. The theorem gives one sufficient way to collapse capacity; it does not characterize all possible regularizers.

G.1 FORWARD POLICY-KL PROOF

The complete proof of Theorem 11 is given in Section 7.

The $\alpha + \eta = 0$ case yields zero penalty and zero occupancy distance for every regularized optimum when $\lambda > 0$. The conclusion concerns candidates that actually meet the regularized optimality condition. Estimating a penalty from finite on-policy samples does not, without uniform error control, establish a bound for every policy in a learned class. A deterministic or shielded reference can fail the support assumption, and adding support changes the reference policy whose trust must be assessed. Reverse-KL penalties and penalties averaged under the reference occupancy are different objectives and are not covered by this proof.

H ADDITIONAL BENCHMARKS

This appendix collects the benchmarks deferred from Section 8: a tabular trigger-room realization of the sparse-signature phenomenon, deterministic noisy-trigger sample scales, exact enumeration under the regularized objective, and model-based continuous- and discrete-action calibrations. As in the main text, these tables contain analytic consequences of declared signal models or values exactly reproduced by Appendix I. The calibration listings regenerate the displayed calculations; learned-family occupancy and access results are reported separately and require their experimental records.

H.1 EXACT IDENTIFICATION AND GRID CONTROLS

The full hidden-branch query sweep follows the optimal Bayes law in Theorem 7. The grid table counts optimal routes and records their local imbalance; its logarithmic route count is not asserted to equal occupancy capacity. Appendix D.1 gives a separate discounted packing proof establishing capacity of the same asymptotic order.

H.2 TRIGGER-ROOM TABULAR BENCHMARK

The hidden-branch construction is intentionally minimal. To show the same sparse-signature phenomenon in a tabular environment closer to gridworld tests, we use a *trigger-room* benchmark: a

Table 5: Exact hidden-branch auditing law. The query count is the minimum integer n such that $(n+1)/M \geq 0.9$.

M	$\log_2 M$	Queries for $\geq 90\%$ success	Certified success $(n+1)/M$
16	4	14	0.938
32	5	28	0.906
64	6	57	0.906
128	7	115	0.906
256	8	230	0.902
512	9	460	0.900
1024	10	921	0.900

Table 6: Gridworld negative control. The number of optimal routes is exponential in N, but the sparse-signature condition fails and exact auditing is linear by path-following.

Grid size	Optimal paths M	$\log_2 M$	Lower bound on $b_{\max}$	Path-following queries
4×4	20	4.32	≥ 10	6
5×5	70	6.13	≥ 35	8
6×6	252	7.98	≥ 126	10
7×7	924	9.85	≥ 462	12
8×8	3432	11.74	≥ 1716	14

common corridor leads to one of M side rooms selected uniformly by the transition kernel, each room contains a length-L loop with identical reward, and policy π_i behaves like the common background policy in every room except room i, where it takes a distinguished safe-but-different action on each loop state. The distinguished action does not alter reward or transition, so all π_i are optimal; however, an exact query to a room state returns a non-background action only for the corresponding π_i. This benchmark has the same 1-sparse signature structure as Theorem 7 but places the hidden branch inside an explicit tabular room layout. Table 7 reports the analytic query law.

Table 7: Trigger-room benchmark. This positive sparse-signature tabular benchmark realizes the same lower-bound law as the hidden-branch construction while using room-like state clusters instead of abstract branches.

Rooms M	Room loop length L	Signature sparsity b	Capacity order	Queries for 90%
16	$16M$	1	$\log_2 M = 4$	14
32	$16M$	1	$\log_2 M = 5$	28
64	$16M$	1	$\log_2 M = 6$	57
128	$16M$	1	$\log_2 M = 7$	115

H.3 Noisy-trigger sample scales

For the Bernoulli noisy-trigger benchmark, let $p_0 = 1/2$ and $p_1 = 1/2 + \rho\Delta$. Assume $0 < \rho\Delta < 1/2$ so that the Bernoulli law is valid and the KL is finite. For small $\rho\Delta$, $\mathrm{KL}(p_0 \| p_1) = \Theta(\rho^2\Delta^2)$. Table 8 reports two deterministic budget scales: the information lower scale $M/(\rho^2\Delta^2)$ from Theorem 10, and a conservative scan scale $5M\log(M)/(\rho^2\Delta^2)$ used by the reproduction script in Appendix I. The scan column is not a theorem lower bound; it reflects the extra multiple-testing burden of a max-scan GLRT at fixed false-positive rate.

H.4 Exact enumeration under the regularized objective

We exact-enumerate all 252 shortest paths in a 6×6 gridworld and use the finite-horizon path occupancy measure over the 10 action steps. Choosing the reference path "Right^5 then Down^5", we

Table 8: Noisy-trigger sample scales. The lower scale is information-theoretic; the scan scale includes a conservative $\log M$ multiple-testing factor.

M	ρ	Δ	Lower scale $M/(\rho^2\Delta^2)$	Conservative scan scale
64	0.1	0.2	160000	3327107
64	0.2	0.2	40000	831777
64	0.2	0.4	10000	207945
256	0.1	0.2	640000	17744568
256	0.2	0.2	160000	4436142
256	0.2	0.4	40000	1109036
1024	0.1	0.2	2560000	88722840
1024	0.2	0.2	640000	22180710
1024	0.2	0.4	160000	5545178

compute

$$\Pi_{\alpha,\lambda} = \left\{\pi \in \Pi_{\text{opt}} : \widetilde{V}_\lambda(\pi) \geq \max_{\pi'} \widetilde{V}_\lambda(\pi') - \alpha\right\}, \qquad \alpha = 0.01.$$

Because all shortest paths have identical unregularized reward, membership depends exactly on whether $\mathrm{TV}(d^\pi, d^0) \leq \alpha/\lambda$. Table 9 reports exact counts reproduced by the Appendix code; this verifies the TV-penalty membership identity on a fully enumerated finite-horizon instance. The remaining log count is not itself a discounted occupancy covering number.

Table 9: Exact enumeration under the canonical regularized objective on a 6×6 gridworld. The counts are obtained by enumerating all 252 shortest paths and measuring finite-horizon path-occupancy TV distance to a reference path.

Penalty λ	Allowed TV α/λ	$\lvert\Pi_{\alpha,\lambda}\rvert$	Remaining $\log_2 \lvert\Pi_{\alpha,\lambda}\rvert$
0.00	∞	252	7.98
0.01	1.00	252	7.98
0.02	0.50	16	4.00
0.03	0.33	4	2.00
0.05	0.20	2	1.00
0.10	0.10	1	0.00

H.5 Continuous-action calibration

For continuous-control calibration, we instantiate the noisy-trigger model with an SAC-style stochastic action head. Before the deterministic tanh/action-range transform, the nominal action is modeled as

$$z \sim \mathcal{N}(\mu, \sigma^2 I_8), \qquad \sigma = 0.5,$$

and a trigger shifts the mean to $\mu + \delta$. We define $\Delta = \|\delta\|_2$, so under the shared isotropic covariance model

$$\beta = \mathrm{KL}\big(\mathcal{N}(\mu, \sigma^2 I_8) \big\| \mathcal{N}(\mu + \delta, \sigma^2 I_8)\big) = \frac{\Delta^2}{2\sigma^2}.$$

Thus $\Delta = 0.1$ and 0.5 correspond to analytic per-trigger signals $\beta = 0.020$ and 0.500 nats. We assign ten candidate trigger locations per path, giving $M = 10P$ for $P \in \{2, 8, 16\}$ paths.

These quantities are a model-based calibration of Theorem 10, not estimates of an unknown SAC noise parameter. No Monte Carlo detector-success claim is required for the lower-bound calculation. Because the same deterministic bijective tanh/action-range transform is applied under both hypotheses, the pre-squash KL is preserved on the transform's image; the displayed formula would not apply unchanged to post-clipping shifts or unequal covariances.

Table 10: Model-based continuous-action calibration of the noisy-trigger bound. The pre-squash action model is $\mathcal{N}(\mu, 0.5^2 I_8)$, $\Delta = \|\delta\|_2$ is the Euclidean mean displacement, and each path is assigned ten candidate trigger locations, so $M = 10P$. Signal KL and sample budgets are analytic consequences of these declared assumptions rather than estimates from detector trials.

Paths	Triggers M	Shift Δ	Analytic signal β	Lower scale M/β	Conservative scan scale
2	20	0.1	0.020	1000	14979
2	20	0.5	0.500	40	600
8	80	0.1	0.020	4000	87641
8	80	0.5	0.500	160	3506
16	160	0.1	0.020	8000	203007
16	160	0.5	0.500	320	8121

H.6 Discrete-action calibration

For the discrete-action calibration, we instantiate an ϵ-greedy policy with $\epsilon = 0.1$ over $A = 6$ actions. Under the nominal policy, the greedy action has probability

$$1 - \epsilon + \frac{\epsilon}{A} = \frac{55}{60},$$

while each alternative has probability $1/60$. A trigger swaps the probability masses of the greedy action and one designated alternative, leaving the remaining four actions unchanged. The resulting forward KL is

$$\beta = \frac{55}{60} \log 55 + \frac{1}{60} \log \frac{1}{55} = 0.9 \log 55 = 3.6066 \text{ nats}.$$

Table 11 maps this declared action signal to the information-theoretic and conservative scan scales for candidate sets of size 100, 1000, and 10000. These are analytic calibration results; no simulated detector-success rates are used in the paper's claims.

Table 11: Model-based discrete-action calibration under ϵ-greedy sampling with $\epsilon = 0.1$ and six actions. The trigger swaps the greedy-action probability with one designated alternative, yielding analytic per-trigger KL $\beta = 0.9 \log 55 = 3.6066$ nats. The displayed budgets are deterministic consequences of this declared signal model.

Candidate screens M	Analytic signal β	Lower scale $\lceil M/\beta \rceil$	Conservative scan scale
100	3.6066	28	639
1000	3.6066	278	9577
10000	3.6066	2773	127688

H.7 Calibration details summary

Table 12 summarizes the declared non-tabular signal models used to connect the noisy-trigger theorem to continuous- and discrete-action RL. These calibrations are analytic and are separate from the learned-family occupancy and access experiments in Section 8.2.

Table 12: Non-tabular calibration settings used to connect the noisy-trigger theory to continuous- and discrete-action RL models. The signal models and budget calculations are analytic; learned-family occupancy experiments are reported separately in Section 8.2.

Item	Continuous-action calibration	Visual-action calibration
RL context	SAC-style continuous action	DQN-style discrete action
Action dimension	8	6
Nominal model	$\mathcal{N}(\mu, 0.5^2 I_8)$	ϵ-greedy, $\epsilon = 0.1$
Trigger	mean shift δ	greedy/alternative mass swap
Signal parameter	$\Delta = \|\delta\|_2$	$\epsilon = 0.1$
Per-trigger KL	$\Delta^2/(2\sigma^2)$	$0.9 \log 55$
Candidate count	$M = 10P$, $P \in \{2, 8, 16\}$	$M \in \{100, 1000, 10000\}$
Output	theory and scan budgets	theory and scan budgets

I Reproducibility Listings

This appendix reproduces the deterministic calculations behind the analytic control and calibration tables. The listings do not reproduce the learned-family training, projected packings, return intervals, probe statistics, or access-study trials. Those analyses require the corresponding rollout logs, checkpoint identifiers, projection records, per-state action distributions, and audit records.

I.1 Hidden-branch and gridworld count tables

```
import math

# Hidden-branch exact law
for M in [16, 32, 64, 128, 256, 512, 1024]:
    n90 = math.ceil(0.9*M - 1)
    print(M, math.log2(M), n90, (n90+1)/M)

# Shortest-path gridworld negative control
for N in range(4, 9):
    M = math.comb(2*(N-1), N-1)
    print(N, M, math.log2(M), math.ceil(M/2), 2*N-2)
```

I.2 Noisy-trigger sample scales

```
import math

for M in [64, 256, 1024]:
    for rho, Delta in [(0.1, 0.2), (0.2, 0.2), (0.2, 0.4)]:
        sig2 = (rho*Delta)**2
        lower = math.ceil(M / sig2)
        scan = math.ceil(5 * M * math.log(M) / sig2)
        print(M, rho, Delta, lower, scan)
```

I.3 Regularizer enumeration on shortest paths

```
import math
from itertools import combinations
from collections import Counter

N = 6
T = 2*(N-1)

def all_shortest_paths(N):
    T = 2*(N-1)
    for downs in combinations(range(T), N-1):
        downs = set(downs)
        yield ['D' if t in downs else 'R' for t in range(T)]

def occupancy(actions):
    x, y = 0, 0
    d = Counter()
    T = len(actions)
    for a in actions:
        d[((x, y), a)] += 1.0 / T
        if a == 'R':
            x += 1
        else:
            y += 1
    return d

def tv(d1, d2):
    keys = set(d1) | set(d2)
    return 0.5 * sum(abs(d1.get(k, 0.0) - d2.get(k, 0.0)) for k in keys)

ref = ['R']*(N-1) + ['D']*(N-1)
```

```
d0 = occupancy(ref)
dists = [tv(occupancy(p), d0) for p in all_shortest_paths(N)]
alpha = 0.01
for lam in [0.0, 0.01, 0.02, 0.03, 0.05, 0.10]:
    threshold = float('inf') if lam == 0.0 else alpha / lam
    count = sum(1 for z in dists if z <= threshold + 1e-12)
    print(lam, threshold, count, math.log2(count))
```

I.4 Analytic non-tabular calibration calculations

```
import math

# Continuous-action calibration assumptions used in
↪  Table~\ref{tab:antmaze}.
sigma = 0.5
for M_paths in [2, 8, 16]:
    for Delta in [0.1, 0.5]:
        kl = (Delta**2) / (2 * sigma**2)
        triggers_per_path = 10
        M = M_paths * triggers_per_path
        lower = math.ceil(M / kl)
        scan = math.ceil(5 * M * math.log(M) / kl)
        print(M_paths, M, Delta, kl, lower, scan)

# Discrete-action calibration assumptions used in
↪  Table~\ref{tab:visual-rl}.
eps = 0.1
A_size = 6
p0 = [1 - eps + eps/A_size] + [eps/A_size]*(A_size-1)
p1 = [eps/A_size] + [1 - eps + eps/A_size] + [eps/A_size]*(A_size-2)

def kl_div(p, q):
    return sum(p[i] * math.log(p[i]/q[i]) for i in range(len(p)))

kl_visual = kl_div(p0, p1)
for M_visual in [100, 1000, 10000]:
    lower = math.ceil(M_visual / kl_visual)
    scan = math.ceil(5 * M_visual * math.log(M_visual) / kl_visual)
    print(M_visual, kl_visual, lower, scan)
```

I.5 Exact root-reset budgets

The recursion tracks the distribution of the number of distinct sampled branches and includes the saturation at full coverage.

```
import math

for M in [64, 128, 256]:
    prob = [0.0] * (M + 1)
    prob[0] = 1.0
    n = 0
    while True:
        success = sum(prob[d] * min(1.0, (d + 1) / M)
                      for d in range(M + 1))
        if success >= 0.9:
            print(M, math.ceil(0.9*M - 1), n, success)
            break
        nxt = [0.0] * (M + 1)
        for d, mass in enumerate(prob):
            nxt[d] += mass * d / M
            if d < M:
                nxt[d + 1] += mass * (M - d) / M
        prob = nxt
        n += 1
```