THE LAST CRITIC BUYS PARALLELISM: A QUERY–FEEDBACK LAW FOR CRITIC LIBRARIES

Anonymous authors
Paper under double-blind review

ABSTRACT

A critic can be redundant for exact identification yet valuable for parallel acquisition. We study n separately optimized singleton-avoidance critics in a finite discounted MDP with n states and A actions per state. Any $n-1$ critics identify the transition kernel exactly, but noise makes the identity of a stable omission depend on the unknown kernel. We ask how many feedback rounds are needed to match the complete library's recovery and execution rates as the horizon and noise level vary, when each batch is chosen before any of its columns arrives. For any fixed row-diameter bound $0 < \kappa < 1$, let $b = \lceil 1/\kappa \rceil$. When $n \geq 3$, a blockwise residual rule for arbitrary finite nonempty action sets uses $n-1$ queries in at most $\lceil (n-2)/(b-1) \rceil$ rounds and achieves the same error scaling. If $n \geq b+2$ and $A \geq b+3$, a hidden-core lower bound shows that at least $\lceil (n-1)/b \rceil$ rounds are necessary. Under the same conditions, if each round can carry at most W critics, the minimum identifying budget has depth $\Theta(1 + n \max\{\kappa, W^{-1}\})$. Without a per-round cap, these depths are $\Theta(1 + n\kappa)$ for $n-1$ critics and one for all n. The penultimate critic buys identifiability; the final buys parallelism by removing the adaptive bottleneck, not the per-round cap.

1 INTRODUCTION

An agent may retain optimal critics for basic tasks after the data used to learn them are gone. Such a library can support new rewards because its values encode properties of the environment. Bellman inversion formalizes this connection: a rich enough value collection identifies transition dynamics, while approximate values recover them with a conditioning-dependent error (Letcher et al., 2026). This raises an acquisition question. If a query budget requires leaving at least one critic unread, which critics should the agent request before the downstream task is known, and how many rounds of feedback does it need?

For each state, we consider an optimal critic that minimizes the expected discounted number of future visits to that state. Any $n-1$ of these n singleton-avoidance critics identify an n-state kernel exactly. It is tempting to view the final critic as useless. Noise changes the problem. A transition perturbation can appear directly in one critic and only through discounted continuation in the others. Which column is safe to omit then depends on the kernel. The observed columns must guide that choice.

The last two critics buy different resources. The penultimate critic makes the dynamics identifiable. The final critic adds no noiseless identifying dimension, but it removes the need to choose a stable omission. For $0 < \kappa < 1$, let $b = \lceil 1/\kappa \rceil$, where κ bounds the transition-row diameter. At the minimum identifying budget, the block rule matches the complete library's error scaling with horizon and noise level using $O(1 + n\kappa)$ feedback rounds for arbitrary finite action sets. When $n \geq b+2$ and $A \geq b+3$, a matching lower bound requires $\Omega(1+n\kappa)$ rounds; complete acquisition needs one.

This law separates query count from feedback depth. A round may request an arbitrary batch, but all labels in it are fixed before any reply arrives. All replies come from one noisy table fixed before the learner's private randomness. The learner receives no transition samples during acquisition and sees the downstream reward only afterward. During evaluation, it may continue learning from its full trajectory.

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.

The upper bound first requests $\min\{b, n-1\}$ arbitrary critics. If a suffix remains, it then requests up to $b-1$ labels with the smallest signed residuals per round. The proof uses an inverse-sign structure for the minimum-occupation matrix and a dimension-free bound on joint centered occupation. Together with the positive margin $1-(b-1)\kappa$, these properties keep one stable label unqueried along the noisy query path. The lower bound hides a core of $b+1$ labels. Packing the queries into fewer rounds can exhaust that core before its replies are observed, leaving the direct measurement unseen. This argument permits randomized schedules, adaptive later rounds, and history-dependent online execution.

Separate task optimization is essential. Under one common continuation policy, a conservation identity completes the missing column, so $n-1$ critics may be requested in one batch. Goal-specific policies remove that common certificate: compatible models can disagree on which redundant coordinate is stable. Appendix M gives the formal comparison.

2 Protocol and feedback depth

Critic library. The known state space S has n states, finite nonempty action sets A_s, and discount $0 \le \gamma < 1$. Put $H = (1-\gamma)^{-1}$. All tasks share an unknown stochastic kernel P. For a bounded arrival reward $r \in \mathbb{R}^n$, define

$$Q^*_{P,r}(s,u) = \sup_{\pi} \mathbb{E}_P \left[\sum_{h \ge 0} \gamma^h r(S_{h+1}) \,\middle|\, S_0 = s, A_0 = u \right]. \tag{1}$$

Write $J^*_P(o;r) = \max_{u \in A_o} Q^*_{P,r}(o,u)$. All returns are continuing: entering a target state does not terminate the trajectory. For a state cost c, write $q_c = -Q^*_{P,-c}$, and let V_c include the current-state cost. Then

$$V_c(s) = c(s) + \gamma \min_{u \in A_s} q_c(s,u), \qquad q_c(s,u) = P_{su} V_c. \tag{2}$$

The singleton-avoidance critic for label g uses $c = e_g$ and is denoted q_g. Each label may induce a different optimal continuation policy. Let q_P be the matrix formed by these n state–action columns, and set $\bar{q}_P = q_P / H$.

Batched acquisition. An instance fixes a complete table Z satisfying

$$\|Z - \bar{q}_P\|_{\max} \le \delta. \tag{3}$$

A query for g returns the entire column Z_g. The table, including all unqueried entries, is fixed before the learner's private randomness; its errors may be arbitrarily correlated. In one feedback round the learner chooses a batch of previously unqueried labels, then receives all their columns. A later batch may depend on every earlier reply. Acquisition uses at most q distinct labels in at most t nonempty rounds. It receives no transition samples and ends before the downstream reward is disclosed. If a jointly learned table satisfies (3) with probability at least $1-\alpha$ before acquisition randomness is drawn, the deterministic bounds below hold on that event; the critic errors may remain correlated.

Define the global transition-row diameter by

$$\kappa(P) := \max_{s,u,s',u'} \mathrm{TV}(P_{su}, P_{s'u'}). \tag{4}$$

We assume $\kappa(P) \le \kappa$ for a fixed $0 < \kappa < 1$. This coefficient controls the contraction of value spans (Gaubert & Qu, 2015). Rate statements use a common action count $|A_s| = A$; additional action labels may duplicate existing rows.

Recovery risk. Let $\mathfrak{A}_q^{(t)}$ be the class of batched acquisition rules just defined. After acquisition, a possibly randomized estimator outputs $\widehat{P}$. The minimax maximum-row error is

$$\mathcal{R}_q^{(t)}(\delta; H, \kappa) = \inf_{\substack{\mathsf{A} \in \mathfrak{A}_q^{(t)} \\ \widehat{P}}} \sup_{\substack{\kappa(P) \le \kappa \\ \|Z - \bar{q}_P\|_{\max} \le \delta}} \mathbb{E}\left[\max_{s,u} \|\widehat{P}_{su} - P_{su}\|_1 \right]. \tag{5}$$

The expectation is over private randomness. Both P and the complete table are chosen before that randomness.

Execution risk. After acquisition, a reward $\rho \in [0,1]^n$ and initial state o are revealed. They are part of the instance fixed before private randomness. An execution rule Π may depend on the acquired columns and on its full trajectory. Define

$$L_P(\Pi;\rho,o) = \frac{J_P^*(o;\rho) - \mathbb{E}_{P,Z,\rho,o}^{\mathsf{A},\Pi}\left[\sum_{h\ge 0}\gamma^h \rho(S_{h+1})\right]}{H}. \tag{6}$$

The expectation includes acquisition randomness, policy randomness, and the trajectory. The corresponding minimax risk is

$$\mathcal{D}_q^{(t)}(\delta;H,\kappa) = \inf_{\substack{\mathsf{A}\in\mathfrak{A}_q^{(t)}\\ \Pi}} \sup_{\substack{\kappa(P)\le\kappa,\ \|Z-\bar{q}_P\|_{\max}\le\delta\\ \rho\in[0,1]^n,\ o\in S}} L_P(\Pi;\rho,o), \tag{7}$$

where A uses at most q queries and t acquisition rounds. No restriction is placed on later environment interactions.

Complete-library precision and feedback depth. For fixed n, A, κ, use the benchmarks

$$G_{\mathrm{rec}}(\delta,H) = \min\{1, nH\delta\}, \qquad G_{\mathrm{exec}}(\delta,H) = \min\left\{1, \delta H, \sqrt{\frac{\log(eH)}{H}}\right\}. \tag{8}$$

The first is the complete library's sharp recovery rate. The second is its execution rate up to the logarithm in the online-learning branch (Fruit et al., 2018). Let $\mathcal{X} = \mathcal{R}$ when $x = \mathrm{rec}$, and let $\mathcal{X} = \mathcal{D}$ when $x = \mathrm{exec}$. Define

$$d_x(q) = \min\left\{t \ge 1 : \begin{array}{l}\exists K_{n,A,\kappa} < \infty \text{ such that}\\ \mathcal{X}_q^{(t)}(\delta;H,\kappa) \le K_{n,A,\kappa} G_x(\delta,H) \quad \forall H \ge 2,\ \delta \ge 0\end{array}\right\}. \tag{9}$$

Set $d_x(q) = \infty$ if the set is empty. The algorithms may use the known problem parameters. The constant $K_{n,A,\kappa}$ may depend on fixed n, A, κ, but not on H or δ. Thus the depth measures how much feedback is needed to preserve the complete library's horizon–noise order.

At $\delta = 0$, one batch of $n-1$ exact columns already identifies the kernel. Feedback depth concerns the stronger guarantee that holds uniformly over all horizons and noise levels.

The minimax infimum may use the declared error budget. Our acquisition and fitting rules do not use it; only the post-acquisition coordinator uses δ to choose between fitted-model planning and online learning.

3 The Query–Feedback Law

The benchmark in (9) is sharp on the same overlap class used by the round comparison.

Proposition 1 (Complete-library benchmark)**.** *For every $n \ge 3$, $A \ge 1$, $0 < \kappa < 1$, $H \ge 2$, and $\delta \ge 0$,*

$$\mathcal{R}_n^{(1)}(\delta;H,\kappa) \asymp_\kappa \min\{1, nH\delta\}. \tag{10}$$

Put $b = \lceil 1/\kappa \rceil$. If $n \ge b+2$ and $A \ge b+3$, then there are constants $c_\kappa > 0$ and $C_{n,A,\kappa} < \infty$ such that

$$c_\kappa \min\{\delta H, H^{-1/2}\} \le \mathcal{D}_n^{(1)}(\delta;H,\kappa) \le C_{n,A,\kappa} G_{\mathrm{exec}}(\delta,H). \tag{11}$$

The recovery lower bound uses two kernels that admit the same complete noisy table. Its constant can decrease as κ approaches zero. The execution lower bound permits history-dependent learning after the reward is disclosed. The logarithmic gap in (11) is the one already present in the online fallback.

The block upper bound below holds for arbitrary finite nonempty action sets. The condition $A \ge b+3$ enters only through the matching hidden-core lower bound in the branch $b \le n-2$.

Theorem 1 (Query–feedback law)**.** *Fix $n \ge 3$ and $0 < \kappa < 1$, put $b = \lceil 1/\kappa \rceil$, and let $x \in \{\mathrm{rec}, \mathrm{exec}\}$.*

If $n \geq b+2$ and $A \geq b+3$, then

$$d_x(q) = \begin{cases} \infty, & 0 \leq q \leq n-2, \\ \text{an integer in } \left[\left\lceil \frac{n-1}{b} \right\rceil, \left\lceil \frac{n-2}{b-1} \right\rceil\right], & q = n-1, \\ 1, & q = n. \end{cases} \tag{12}$$

If $b \geq n-1$, then $d_x(n-1) = 1$, with no lower bound on the number of actions. Consequently, over the full overlap range,

$$d_x(n-1) = \Theta(1 + n\kappa), \tag{13}$$

where the round-count constants are universal and $A \geq b+3$ is assumed in the branch $b \leq n-2$.

Corollary 1 (Capacity-constrained feedback)**.** *For $x \in \{\mathrm{rec}, \mathrm{exec}\}$, let the integer $W \geq 1$ bound the number of critics per round. Define $d_{x,W}$ as in* (9)*, restricting every batch to at most W critics. If $n \geq b+2$ and $A \geq b+3$, then*

$$d_{x,W}(n-1) = \Theta\big(1 + n \max\{\kappa, W^{-1}\}\big), \tag{14}$$

with universal comparison constants.

The capacity and hidden-core arguments give the lower bounds. For $W = 1$, split the width-two rule's initial pair across two rounds; it then uses $n-1$ rounds. For $W \geq 2$, the tunable-width rule proved in Section 5 and its execution transfer use $m = \min\{W, b\}$. Capacity dominates until it reaches the largest uniformly stable block width; feedback dominates afterward.

Exact one-round boundary. Under the theorem's lower-bound action condition, the minimum identifying budget has depth one if and only if $(n-2)\kappa < 1$.

Table 1: **The last two critics buy different resources.** The table shows the regime $n \geq b+2$, $A \geq b+3$ in Theorem 1. The middle depth is determined up to universal constant factors.

Queries	What remains unresolved	Feedback depth	Resource obtained
$q \leq n-2$	An identifying dimension	∞	Insufficient for exact identification
$q = n-1$	Which critic is stable to omit	$\Theta(1+n\kappa)$	Identifiability
$q = n$	No omission decision	1	Parallel acquisition

The first transition is algebraic. With two critics omitted, two kernels can produce the same exact adaptive transcript with positive probability. The complete-library benchmark is zero at $\delta = 0$, so further critic feedback cannot close the missing identifying dimension.

The middle transition is about robust selection. For the lower bound, packing $n-1$ queries into fewer than $\lceil (n-1)/b \rceil$ rounds forces some batch to contain at least $b+1$ labels. A hidden core can occupy that batch and consume the safe omission before its replies are available. For the upper bound, request any b columns first and then at most $b-1$ labels with the smallest signed residuals in each later round. The resulting number of rounds is $\lceil (n-2)/(b-1) \rceil$. The acquisition rule uses the known κ, but neither δ nor the downstream reward.

The final transition removes selection. With all n columns, the learner does not need to decide which coordinate is stably redundant, so one batch suffices. Equation (13) also covers strong overlap: when $b \geq n-1$, all $n-1$ identifying columns may be requested safely in the first round.

4 Exact redundancy and stable omission

Form the current-state minimum-occupation matrix

$$M_{sg} = \mathbf{1}\{s = g\} + \gamma \min_{a \in A_s} q_g(s, a). \tag{15}$$

Its first-step identity is

$$q(s, a, \cdot) = P_{sa} M. \tag{16}$$

Each column may use a different optimal policy. Nevertheless, controlled state elimination gives the columns a joint inverse structure.

Theorem 2 (Minimum-occupation structure)**.** *Let each state have a finite nonempty set of nonnegative action rows $K_{s,a}$ with $K_{s,a}\mathbf{1} \le \gamma < 1$. Let column g of M solve*

$$M_{sg} = \mathbf{1}\{s = g\} + \min_a K_{s,a} M_{\cdot g}. \tag{17}$$

Then M is invertible and $C = M^{-1}$ satisfies

$$0 < C_{ii} \le 1, \quad C_{ij} \le 0\ (i \ne j), \quad C\mathbf{1} \ge (1-\gamma)\mathbf{1}, \quad \|C\|_\infty \le 1 + \gamma. \tag{18}$$

The matrix norm is the maximum absolute row sum; its bound is sharp.

The MDP case has $K_{s,a} = \gamma P_{sa}$. Allowing discounted killing keeps this class closed under state elimination: removing one state produces another minimum-occupation problem with an enlarged action family. The proof then proceeds by induction. A lower-envelope lemma and a positive Schur complement give the inverse signs; a Bellman residual bounds the diagonal and row sums. Appendix J gives the full proof and its relation to classical matrix methods.

Writing $L = I - C$ gives $L \ge 0$, $L\mathbf{1} \le \gamma\mathbf{1}$ and $M = (I - L)^{-1}$. Thus the separately optimized columns form the Green matrix of an effective process that may terminate. It need not be a policy of the original MDP. The common controlled action family is essential; arbitrary entrywise minima of Green matrices lack this closure (Appendix S).

Which omitted coordinate is stable? Put $c = C\mathbf{1}$ and call its entries c_g the inverse masses. Theorem 2 gives $H^{-1} \le c_j \le 1$. For a probability-row difference ξ, let $y = \xi M$ and write C_g for row g of C. Then $yc = \xi\mathbf{1} = 0$. Omitting label j yields

$$\xi = \sum_{g \ne j} y_g \left(C_g - \frac{c_g}{c_j} C_j \right). \tag{19}$$

If every observed y_g is zero, then $\xi = 0$. This proves that any $n - 1$ singleton-avoidance columns identify the kernel. With noise, however, the same expression may contain a factor c_g/c_j as large as H. Exact redundancy alone does not control amplification.

Writing $S_c = \sum_g c_g$, the same identity gives

$$\|\xi\|_1 \le 2 \left[(n-2) + \frac{S_c}{c_j} \right] \|\xi M_{S\setminus\{j\}}\|_\infty, \qquad \xi\mathbf{1} = 0. \tag{20}$$

Thus, for any fixed $\alpha > 0$, leaving a label j unqueried with $c_j \ge \alpha S_c/n$ is sufficient for linear amplification in n. Computing c directly would require the complete matrix. The acquisition rule must protect such a coordinate while leaving its column unseen.

5 Feedback depth at the minimum query budget

Exact identifiability does not determine how often acquisition must use feedback. Any $n - 1$ singleton-avoidance critics identify the transition kernel, but exact redundancy alone does not guarantee stable decoding under noise. We show that the minimum query budget retains the complete library's $nH\delta$ recovery order while acquiring several critics between responses.

From algebraic redundancy to stable redundancy. Let M be the minimum-occupation matrix, $C = M^{-1}$, $c = C\mathbf{1}$, and $S_c = \mathbf{1}^\top c$. Equation (20) gives a sufficient condition for stable decoding: leave a label j unqueried with $c_j \ge \alpha_\kappa S_c/n$, where $\alpha_\kappa > 0$ may depend on κ, but not on n, H, δ. The small-noise analysis below establishes this property using only returned columns; for larger noise, the probability-row cap gives the uniform bound. The vector c depends on unseen columns, so it cannot be computed directly during acquisition.

Fix $0 < \kappa < 1$, assume $\kappa(P) \le \kappa$, and put

$$b = \left\lceil \frac{1}{\kappa} \right\rceil, \qquad d = 1 - (b-1)\kappa > 0. \tag{21}$$

The integer b controls the first batch. Later batches contain at most $b - 1$ labels, the largest uniform block size for which the overlap margin d is strictly positive.

Block acquisition. The rule begins with any prescribed set I_0 of $\min\{b, n-1\}$ labels. Given the acquired set I, let $U = S \setminus I$. If $\widetilde{M}_{II}$ is nonsingular, form

$$\widehat{x} = (\widetilde{M}_{II})^{-1}\mathbf{1}, \qquad \widetilde{s}_U = \mathbf{1} - \widetilde{M}_{UI}\widehat{x}, \tag{22}$$

where $\widetilde{M}_{sg} = \mathbf{1}\{s = g\} + \gamma H \min_a Z_g(s, a)$. The acquired columns interpolate the constant vector at queried states: $\widetilde{M}_{II}\widehat{x} = \mathbf{1}$. For an unqueried label g, the signed residual $\widetilde{s}_g$ is the remaining interpolation error at state g. Although column g is unqueried, every matrix entry in its score belongs to a column already in I. The rule then queries, in one batch, the $\min\{b-1, |U|-1\}$ smallest signed residuals in $\widetilde{s}_U$, with fixed ties. If $\widetilde{M}_{II}$ is singular, it instead queries a fixed unused block of the same size. The procedure orders signed values, not their absolute magnitudes, and stops after $n-1$ columns.

In the exact exclusion-core construction of Section 6, after b core queries the last core residual is 1, while every outside residual is $1 - \gamma b/(b+1)$. The rule therefore queries outside the core first and leaves a high-inverse-mass core label unread.

Let J be the final acquired set. For each state-action pair, the rule solves

$$\widehat{P}_{sa} \in \arg\min_{p \geq 0,\ p\mathbf{1}=1} \|p\widetilde{M}_J - HZ_J(s, a, \cdot)\|_\infty, \qquad |J| = n - 1. \tag{23}$$

Selection and fitting use the same received table. They know κ, through b, but use neither the noise budget δ nor the downstream reward.

Theorem 3 (Block feedback upper bound)**.** *Let $n \geq 3$, $H \geq 1$, $0 < \kappa < 1$, and $\kappa(P) \leq \kappa$. For finite nonempty action sets and every complete table satisfying $\|Z - q_P/H\|_{\max} \leq \delta$, the block rule queries $n-1$ columns in at most*

$$r_{\mathrm{fb}}(n, \kappa) \leq \left\lceil \frac{n-2}{b-1} \right\rceil, \qquad b = \lceil 1/\kappa \rceil, \tag{24}$$

rounds and returns a stochastic kernel satisfying

$$\max_{s,a} \|\widehat{P}_{sa} - P_{sa}\|_1 \leq \min\{2, C_\kappa n H \delta\}. \tag{25}$$

Here $C_\kappa < \infty$ depends on κ, but not on n, H, δ, or the action counts. When $b \geq n-1$, the rule uses one batch. Otherwise it uses one batch of b columns followed by batches of at most $b-1$, attaining the round count in (24)*.*

Corollary 2 (Tunable block width)**.** *Fix $2 \leq m \leq n-1$ with $(m-1)\kappa < 1$. Query any m labels first, then at most $m-1$ smallest-residual labels per round. This rule queries $n-1$ columns and, for a finite $C_{\kappa,m}$, satisfies*

$$r_{\mathrm{fb}}(n; m) \leq \left\lceil \frac{n-2}{m-1} \right\rceil, \qquad \max_{s,a} \|\widehat{P}_{sa} - P_{sa}\|_1 \leq \min\{2, C_{\kappa,m} n H \delta\}. \tag{26}$$

The constant is independent of n, H, δ, and the action counts. Once m is fixed, acquisition and fitting use neither κ nor δ.

Set $\kappa_m = \max\{\kappa, 1/m\}$. Since $(m-1)\kappa < 1$, $\kappa(P) \leq \kappa \leq \kappa_m$ and $m - 1 < 1/\kappa_m \leq m$, so $\lceil 1/\kappa_m \rceil = m$. Theorem 3 applied with κ_m gives the claim. Thus $m = 2$ gives a κ-oblivious rule with $n-2$ rounds on each fixed class $\kappa < 1$, although its constant need not be uniform as $\kappa \uparrow 1$.

Thus the minimum identifying budget preserves the complete library's horizon–noise order in $O(1 + n\kappa)$ feedback rounds. Section 6 shows that this order is necessary.

Why blocks remain stable. The proof combines one global and one local fact. Center every column at its minimum:

$$m_g = \min_s M_{sg}, \qquad D = M - \mathbf{1}m^\top. \tag{27}$$

Although different columns use different optimal policies, their centered occupations obey a dimension-free joint budget

$$\max_s \sum_g D_{sg} \leq L_\kappa < \infty. \tag{28}$$

Here L_κ depends only on the strict contraction coefficient; for $\kappa \le 1/2$, one may take $L_\kappa < 32$. This is the total amount that the proof charges over the acquisition path.

For the local step, consider a current unqueried set U, the next block $B \subset U$, and $V = U \setminus B$. Put $u = C_{VV}^{-1} c_V$ and $Q = -C_{VV}^{-1} C_{VB}$. If s_U is the exact residual before the block, block inversion gives

$$s_V = u + Q s_B, \qquad Q \ge 0, \qquad Q\mathbf{1} \le \mathbf{1}. \tag{29}$$

Writing $\Delta_h = D_{hh}$, the centered inverse identity sharpens every entry to

$$Q_{jh} \le \gamma\kappa + u_j \frac{m_h}{\Delta_h}, \qquad j \in V,\ h \in B. \tag{30}$$

The direct contribution of a block is at most $(b-1)\kappa = 1-d$. The remaining term is charged across inverse-mass scales. If the final omission had too little inverse mass, noisy selection would incur a fixed charge at every scale. Yet the scale integral charges each column ℓ by at most $D_{j\ell}$; summing over ℓ contradicts the joint occupation bound (28). Appendix A gives the complete perturbation and decoding argument.

Transfer to execution. If the maximum transition-row error is e, every true-policy value has span at most $(1-\gamma\kappa)^{-1} \le (1-\kappa)^{-1}$. Comparing the true optimum with an optimal policy in the fitted model gives normalized loss at most $e/(1-\kappa)$; the fitted model itself need not contract. Consequently the block rule has

$$\mathrm{loss}_{\mathrm{exec}} \le \min\{1, C_\kappa n H \delta/(1-\kappa)\}. \tag{31}$$

If the declared noise budget is available after acquisition, an execution coordinator may instead use the critic-free online rule, yielding $\min\{1, C_\kappa n H\delta/(1-\kappa), B_\kappa(H)\}$, where $B_\kappa(H) = O_{n,A,\kappa}(\sqrt{\log(eH)/H})$. The coordinator uses δ only to select between two already valid branches.

6 A HIDDEN CORE FORCES FEEDBACK

We now show that feedback rounds are a genuine resource, even when the number of acquired critics is already sufficient for exact identification. Write $\mathcal{R}_q^{(r)}$ for the minimax maximum-row transition-recovery risk after at most q queries in at most r feedback rounds, and write $\mathcal{D}_q^{(r)}$ for the corresponding normalized execution risk. The latter permits arbitrary history-dependent learning during the evaluated trajectory. Appendix B gives the formal protocol and all proofs.

Fix $0 < \kappa < 1$ and put

$$b = \left\lceil \frac{1}{\kappa} \right\rceil, \qquad \ell = b + 1. \tag{32}$$

For $r \ge 1$, define

$$\Psi_{n,r,\ell} = \min_{\substack{k_1+\cdots+k_r = n-1 \\ k_t \in \mathbb{N}_0}} \sum_{t=1}^{r} \binom{k_t}{\ell}, \qquad p_{n,r,\ell} = \frac{\Psi_{n,r,\ell}}{\binom{n}{\ell}(n-\ell)}. \tag{33}$$

The quantity $\Psi_{n,r,\ell}$ counts the fewest hidden cores that must be fully contained in one batch, over all ways of distributing $n-1$ queries among r rounds.

Theorem 4 (Finite-round lower bound)**.** *Suppose $n \ge b+2$, $A \ge b+3$, and $H \ge 2$. For every $r \ge 1$ and $\delta \ge 0$,*

$$\mathcal{D}_{n-1}^{(r)}(\delta; H, \kappa) \ge \frac{p_{n,r,\ell}}{5} \min\left\{ \ell\delta H^2, \frac{1}{32 b^2 \sqrt{H}} \right\}, \tag{34}$$

$$\mathcal{R}_{n-1}^{(r)}(\delta; H, \kappa) \ge 2 p_{n,r,\ell} \min\left\{ \ell\delta H^2, \frac{1}{8b^2} \right\}. \tag{35}$$

The instance prior is fixed before the learner's private randomness. The execution bound allows the learner to update its policy from every state, action, and reward observed after acquisition.

The construction hides an ℓ-label exclusion core R and one special label $j \notin R$. The returned noisy core columns are common midpoints. They reveal neither which probe action is better nor the identity of j. If a batch contains all of R, it was necessarily committed before any of those replies arrived. The remaining query budget can then leave j unread. On this event, the kernels induce the same acquisition transcript and conditional law of the learner's private randomness. In the noise-limited regime, their transition rows differ by an amount of order δH^2, and their optimal probe actions are exchanged. This is why the same event yields both (34) and (35).

The constructed kernels have transition-row diameter exactly $1/b \le \kappa$. In particular, when $\kappa > 1/2$, one has $b = 2$ and the same construction uses a three-label core of diameter $1/2$ inside the larger contraction class.

Balancing the batch sizes minimizes (33). If $n - 1 = rk + s$ with $0 \le s < r$, then

$$\Psi_{n,r,\ell} = (r - s)\binom{k}{\ell} + s\binom{k+1}{\ell}, \qquad p_{n,r,\ell} > 0 \quad \Longleftrightarrow \quad rb < n - 1. \tag{36}$$

Thus the obstruction survives every randomized schedule and adaptive stopping rule whenever fewer than $\lceil (n-1)/b \rceil$ rounds are available.

Corollary 3 (The b versus $b + 1$ boundary)**.** *Under the conditions of Theorem 4, in particular $n \ge b + 2$ and $A \ge b + 3$, matching the complete library's horizon–noise order uniformly over H and δ requires at least*

$$r \ge \left\lceil \frac{n-1}{b} \right\rceil. \tag{37}$$

The combinatorial obstruction first becomes possible at batch size $b+1$. The upper-bound algorithm uses b labels initially and at most $b - 1$ labels in each later round. Appendix B, Eq. (89)*, gives the total-round coefficient. The explicit initial-batch coefficient appears in Appendix G.4, Eqs.* (212)–(213)*.*

To see (37), take $\delta = H^{-3}$ in the execution bound, or $\delta = H^{-2}$ in the recovery bound, and let H grow while n, A, κ, r remain fixed. When $rb < n-1$, the lower bound exceeds every fixed multiple of the complete library's δH precision. The result identifies the number of rounds needed for a guarantee uniform over horizon and noise.

There is a separate obstruction below the identifying budget.

Proposition 2 (Two omitted critics leave exact ambiguity)**.** *Fix $n \ge 3$, $0 < \kappa < 1$, $A \ge 2$, $H \ge 2$, $0 \le q \le n - 2$, and $r \ge 1$. For every acquisition rule using at most q labels in at most r rounds, there are two distinct kernels in the diameter-κ class whose exact critic tables yield identical acquired transcripts on an event of positive probability. Hence*

$$\mathcal{R}_q^{(r)}(0; H, \kappa) > 0, \qquad \mathcal{D}_q^{(r)}(0; H, \kappa) > 0. \tag{38}$$

No number of feedback rounds can compensate for the missing identifying dimension.

The proof chooses a hidden pair of labels and constructs two kernels by swapping two state-independent transition rows. Every exact column outside that pair is the same constant column. An adaptive acquisition rule using at most $n - 2$ labels therefore misses both hidden labels with positive probability, regardless of how many rounds it uses. Conditioning on that event leaves a valid two-model recovery and execution test.

The finite-round theorem and the exact-ambiguity proposition supply the two lower regimes of Theorem 1: robust selection costs feedback at $n - 1$ queries, while exact identification fails below that budget.

7 Four-action recovery outside the hidden core

The block recovery guarantee allows arbitrary finite action sets, whereas the hidden-core lower bound needs $A \ge b + 3$. To test recovery outside that construction, we prespecified an $n = 8$, $A = 4$ study on 64 independent kernels at $\kappa \in \{1/3, 1/4\}$: a zero-common-mass punctured family and a local-with-teleportation common-mass control. Paired horizons and noise give 768 primary low-noise conditions among 1,344 total.

With the same initial batch, residual rule, and seven-column budget, block feedback uses three versus five rounds at $\kappa = 1/3$ and two versus four at $\kappa = 1/4$. Every family–diameter setting met the recovery criterion: in 63 of 64 kernels, at least 90% of primary conditions had block error at most twice the paired one-column error. Within each setting, the median across primary conditions of the maximum row error divided by $nH\delta$ was 1.31–1.59, with 90th percentiles 2.43–2.93 and no visible horizon trend. The certificates in Corollary 4 covered all 768 primary conditions and were at most 0.5; stratum median certificate/error ratios were 1.29–9.83, with largest 90th percentile 12.27. The prespecified control-loss criterion held in three of four settings, but block feedback did not consistently beat fixed omissions in the punctured family. Appendix C gives the protocol and data.

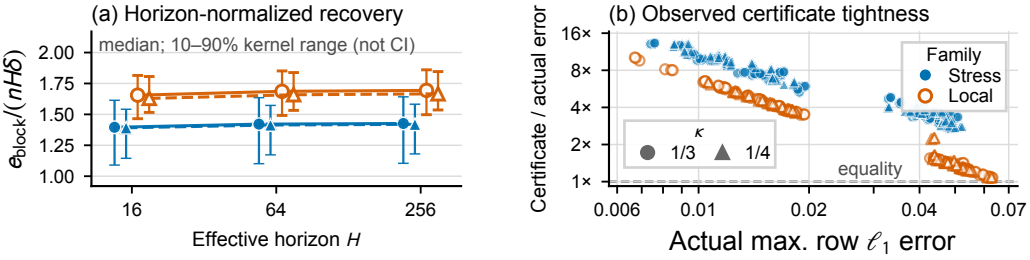

Figure 1: **Four-action recovery with fewer feedback rounds.** (a) Normalized error after collapsing four primary repeats within each kernel; bars show the 10th–90th percentiles over 16 kernels. (b) Certificate/error ratios for 384 kernel–horizon–noise summaries; the dashed line marks equality. Ranges are descriptive, not confidence intervals.

8 Related Work and Scope

Bellman inversion recovers dynamics from values (Letcher et al., 2026), while value equivalence asks which model constraints preserve planning values (Grimm et al., 2020; 2021; 2022). We study feedback when exact-identifying critic subsets differ under noise. Successor features reuse policy-specific occupancies through GPI (Barreto et al., 2017); our separately optimized task columns instead form a joint inverse-M-matrix.

Adaptive matrix sampling uses earlier observations to choose entries or columns (Krishnamurthy & Singh, 2014; Wang & Singh, 2018; Paul et al., 2015); we commit to whole unseen columns. DEIM selects from supplied basis vectors (Sorensen & Embree, 2016); our rule scores each unqueried label using only received columns. Limited-adaptivity and multi-batch RL count sample or policy-update rounds (Agarwal et al., 2017; Zhang et al., 2022); ours counts feedback rounds for pre-reward critic acquisition.

Scope. The interface assumes known finite states, whole-column critic access, no transition samples during acquisition, and a downstream reward revealed afterward. Selection uses neither reward nor noise budget. Blocks wider than two use the known row-diameter bound; width two is κ-oblivious. Acquisition, fitting, and model planning are efficient; the online fallback assumes global confidence-set optimization. Upper bounds cover every $0 < \kappa < 1$ and finite action set; the matching hidden-core lower bound requires $A \geq b + 3$.

9 Conclusion

At the minimum identifying budget, feedback chooses a stable omission without losing complete-library scaling. If $n \geq b+2$, $A \geq b+3$, and each round carries at most $W \geq 1$ critics, the matching depth is $\Theta(1 + n \max\{\kappa, W^{-1}\})$. The final critic adds no identifying dimension: it removes the adaptive bottleneck, not the throughput cap.

REPRODUCIBILITY STATEMENT

The appendices give complete proofs of the structural, recovery, feedback depth, and execution results, including the fixed-prior conditioning used for randomized acquisition rules. The supplement records the prespecified four-action protocol, complete result tables, figure data, generators, and independent replication code. It also includes full traces and fitted kernels for the construction-specific mechanism checks. The minimax theorems are established analytically.

AI USE STATEMENT

Generative AI assisted with hypothesis development, mathematical formulation, proof writing and checking, literature verification, experimental design, construction of synthetic MDP instances, code implementation, data interpretation, figures, and manuscript preparation. The authors retain responsibility for the correctness, attribution, and integrity of the work.

REFERENCES

Arpit Agarwal, Shivani Agarwal, Sepehr Assadi, and Sanjeev Khanna. Learning with limited rounds of adaptivity: Coin tossing, multi-armed bandits, and ranking from pairwise comparisons. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pp. 39–75. PMLR, 2017. URL https://proceedings.mlr.press/v65/agarwal17c.html.

Eitan Altman and Arie Hordijk. Applications of Borovkov's renovation theory to non-stationary stochastic recursive sequences and their control. *Advances in Applied Probability*, 29(2):388–413, 1997. doi: 10.2307/1428009.

K. B. Athreya and P. Ney. A new approach to the limit theory of recurrent Markov chains. *Transactions of the American Mathematical Society*, 245:493–501, 1978. doi: 10.1090/S0002-9947-1978-0511425-0.

André Barreto, Will Dabney, Rémi Munos, Jonathan J. Hunt, Tom Schaul, Hado P. van Hasselt, and David Silver. Successor features for transfer in reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/350db081a661525235354dd3e19b8c05-Abstract.html.

Claude Dellacherie, Servet Martínez, and Jaime San Martín. Inverse M-matrix, a new characterization. *Linear Algebra and its Applications*, 595:182–191, 2020. doi: 10.1016/j.laa.2020.02.024. URL https://doi.org/10.1016/j.laa.2020.02.024.

Ronan Fruit, Matteo Pirotta, Alessandro Lazaric, and Ronald Ortner. Efficient bias-span-constrained exploration-exploitation in reinforcement learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1578–1586. PMLR, 2018. URL https://proceedings.mlr.press/v80/fruit18a.html.

Stéphane Gaubert and Zheng Qu. Dobrushin's ergodicity coefficient for Markov operators on cones. *Integral Equations and Operator Theory*, 81(1):127–150, 2015. doi: 10.1007/s00020-014-2193-2.

Christopher Grimm, André Barreto, Satinder Singh, and David Silver. The value equivalence principle for model-based reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5541–5552. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/3bb585ea00014b0e3ebe4c6dd165a358-Paper.pdf.

Christopher Grimm, Andre Barreto, Greg Farquhar, David Silver, and Satinder Singh. Proper value equivalence. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 7773–7786. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/400e5e6a7ce0c754f281525fae75a873-Paper.pdf.

Christopher Grimm, Andre Barreto, and Satinder Singh. Approximate value equivalence. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 33029–33040. Curran Associates, Inc., 2022. doi: 10.52202/068431-2393. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/d53538ba21c05fa361d2b21704172753-Paper-Conference.pdf.

Arie Hordijk. *Dynamic Programming and Markov Potential Theory*. Mathematisch Centrum, Amsterdam, 1974. URL https://ir.cwi.nl/pub/32117. Doctoral dissertation, Universiteit van Amsterdam.

Akshay Krishnamurthy and Aarti Singh. On the power of adaptivity in matrix completion and approximation. arXiv preprint arXiv:1407.3619v1, 2014. URL https://arxiv.org/abs/1407.3619v1.

Tor Lattimore and Csaba Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. doi: 10.1017/9781108571401.

Dongmin Lee, William Lu, Anuran Makur, and Japneet Singh. Doeblin curves. *IEEE Transactions on Information Theory*, 72(6):3556–3596, 2026. doi: 10.1109/TIT.2026.3678229.

Alistair Letcher, Mattie Fellows, Alexander D. Goldie, Jonathan Richens, Jakob N. Foerster, and Oliver Richardson. Inverting the Bellman equation: From Q-values to world models. arXiv preprint arXiv:2606.21173v1, 2026. URL https://arxiv.org/abs/2606.21173v1.

Jorge J. Moré. A class of nonlinear functions and the convergence of Gauss–Seidel and Newton–Gauss–Seidel iterations. Technical Report 70-123, Computer Science Center, University of Maryland, College Park, Maryland, July 1970. URL https://ntrs.nasa.gov/citations/19710006532. Also catalogued as NASA-CR-116129.

E. Nummelin. A splitting technique for Harris recurrent Markov chains. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 43(4):309–318, 1978. doi: 10.1007/BF00534764.

Saurabh Paul, Malik Magdon-Ismail, and Petros Drineas. Column selection via adaptive sampling. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/d395771085aab05244a4fb8fd91bf4ee-Paper.pdf.

Jeffrey S. Rosenthal. Convergence rates for Markov chains. *SIAM Review*, 37(3):387–405, 1995. doi: 10.1137/1037083.

D. C. Sorensen and Mark Embree. A DEIM induced CUR factorization. *SIAM Journal on Scientific Computing*, 38(3):A1454–A1482, 2016. doi: 10.1137/140978430. URL https://epubs.siam.org/doi/10.1137/140978430.

Arie Tamir. Minimality and complementarity properties associated with Z-functions and M-functions. *Mathematical Programming*, 7:17–31, 1974. doi: 10.1007/BF01585501. URL https://www.math.tau.ac.il/~atamir/minimality_74.pdf.

Yining Wang and Aarti Singh. Provably correct algorithms for matrix column subset selection with selectively sampled data. *Journal of Machine Learning Research*, 18(156):1–42, 2018. URL https://jmlr.org/papers/v18/15-233.html.

Zihan Zhang, Yuhang Jiang, Yuan Zhou, and Xiangyang Ji. Near-optimal regret bounds for multi-batch reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1785. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9bcd1fa0c05e5f25ba7a1261f1852e82-Abstract-Conference.html.

GUIDE TO THE APPENDICES

Appendix A proves the blockwise upper bound, including the noisy block comparison, the cross-scale charge, and the recovery-to-control transfer. Appendix B proves the finite-round hidden-core lower bound, the exact-ambiguity result below the identifying budget, and the complete-library benchmarks used to define feedback depth. Appendix C proves the observed-data certificate, reports the four-action recovery checks, and describes the construction-specific hidden-core mechanism check and independent numerical replication.

The upper-bound proofs combine positive-matrix control (Moré, 1970; Hordijk, 1974; Tamir, 1974; Dellacherie et al., 2020) with regeneration and coupling (Nummelin, 1978; Athreya & Ney, 1978; Rosenthal, 1995; Altman & Hordijk, 1997). Delayed charging handles the distinct optimal policies without a task-count factor.

Appendix D describes the one-column rule and states its structural ingredients; Appendix E proves them. Appendix F states the diameter-$1/2$ recovery and initial-commitment results; Appendix G proves the corresponding execution bounds and the initial boundary. Appendices H and I give complementary cross-scale residual checks and evaluate the observed certificate on their fitted kernels.

Appendix J proves the minimum-occupation inverse structure. The common-mass refinement and its execution proof appear in Appendices K and L; the latter also records the online-learning budget. Appendices M, N, O, and P provide fixed-library recovery, policy-value coordinates, planning bounds, and task-count results. Appendix P.4 gives the fixed-omission construction and its sharp lower bound.

Appendix Q gives a complete-acquisition compression procedure and its recovery and control guarantees. Appendices R and S give the general arrival-reward reset identity and the matrix counterexample. A retained task set may be written I, J, or $\mathcal{I}$ locally; J_P^π denotes a return.

A PROOF OF THE BLOCK FEEDBACK UPPER BOUND

We prove Theorem 3. The argument treats every feedback round as one block; it does not insert an artificial order among the labels queried together.

A.1 PRELIMINARIES

Write $a = 1-\gamma = H^{-1}$, $C = M^{-1}$, $c = C\mathbf{1}$, $S_c = \mathbf{1}^\top c$, and $c_* = \max_g c_g$. Let

$$m_g = \min_s M_{sg}, \qquad D = M - \mathbf{1}m^\top, \qquad \Delta_g = D_{gg}. \tag{39}$$

For $0 < \kappa < 1$, define

$$\begin{aligned} K_\kappa &= \min\left\{k \ge 1 : k\kappa^{k-1} \le \frac{1-\kappa}{4}\right\}, \\ \beta_\kappa &= \frac{3+\kappa}{4}, \qquad \eta_\kappa = \frac{3(1-\kappa)}{4}, \\ L_\kappa &= \frac{K_\kappa}{\beta_\kappa}\left(1 + \frac{1}{\eta_\kappa}\right). \end{aligned} \tag{40}$$

The inverse structure and the parameterized centered-occupation bound in Lemma 1 give

$$\begin{aligned} &0 < C_{ii} \le 1, \quad C_{ij} \le 0\ (i \ne j), \quad a\mathbf{1} \le c \le \mathbf{1}, \quad \|C\|_\infty \le 2, \\ &CD = I - cm^\top, \qquad \max_s \sum_g D_{sg} \le L, \qquad L := L_\kappa < \infty. \end{aligned} \tag{41}$$

When $\kappa \le 1/2$, the six-step specialization also gives the absolute bound $2016/65 < 32$. For row diameter at most κ, the column-shape argument gives

$$0 \le D_{sg} \le \gamma\kappa\Delta_g \quad (s \ne g), \qquad \frac{1}{1+\gamma\kappa} \le \Delta_g \le \frac{1}{1-\gamma\kappa}. \tag{42}$$

In particular, $\Delta_g \ge 1/2$.

We also use the nonnegative row $m^\top C$. If $t_f = 1 - m^\top c$, then $Dc = t_f \mathbf{1}$ and $t_f \geq \Delta_g c_g$. Combining this fact with (41) yields

$$\sum_g m_g \leq \frac{2L}{c_*}. \tag{43}$$

Let $b = \lceil 1/\kappa \rceil$ and

$$d = 1 - (b-1)\kappa > 0. \tag{44}$$

Assume first that $n \geq b + 2$, so the first batch has size b and a nonempty suffix remains. The inverse-mass bound for an arbitrary initial batch applies with

$$\eta = \eta_{\kappa,b} = \min\left\{\frac{1}{2b+1}, \frac{d(1-\kappa)}{6(b+1)}\right\}, \qquad \sum_{g \notin I_0} c_g \geq \eta S_c. \tag{45}$$

Define constants used only in the proof:

$$\alpha = \exp(-4L/d)\min\{\eta, (32L)^{-1}\}, \qquad \varepsilon_0 = \frac{d\alpha}{32n}. \tag{46}$$

A.2 Block interpolation identities

Fix a stage after the first batch. Let I be the acquired set, $U = S \setminus I$, let $B \subset U$ be the next queried block, and put $V = U \setminus B$. The exact interpolation residual is

$$x = M_{II}^{-1}\mathbf{1}, \qquad s_U = \mathbf{1} - M_{UI}x = C_{UU}^{-1}c_U. \tag{47}$$

Put

$$A = C_{VV}^{-1}, \qquad u = Ac_V, \qquad Q = -AC_{VB}. \tag{48}$$

Principal inverse-M-matrix structure gives $A, Q \geq 0$. Partitioning $C_{UU}s_U = c_U$ gives

$$s_V = u + Qs_B. \tag{49}$$

Moreover,

$$\mathbf{1} - Q\mathbf{1} = A(C_{VV}\mathbf{1} + C_{VB}\mathbf{1}) = A(c_V - C_{VI}\mathbf{1}) \geq 0,$$

so $Q\mathbf{1} \leq \mathbf{1}$.

Restrict $CD = I - cm^\top$ to rows V and columns B, multiply by A, and use $AC_{VI} \leq 0$ and $D_{IB} \geq 0$. This gives

$$QD_{BB} = D_{VB} + AC_{VI}D_{IB} + um_B^\top \leq D_{VB} + um_B^\top. \tag{50}$$

Taking row $j \in V$, column $h \in B$, and retaining only the nonnegative diagonal term on the left yields

$$Q_{jh} \leq \gamma\kappa + u_j \frac{m_h}{\Delta_h}. \tag{51}$$

Finally, (43) and $\Delta_h \geq 1/2$ imply

$$\sum_{h \in B} \frac{m_h}{\Delta_h} \leq \frac{4L}{c_*}. \tag{52}$$

A.3 A stable label survives every block

Proposition 3 (Blockwise stable omission). *If $\|\widetilde{M} - M\|_{\max} \leq \varepsilon_0$, the block acquisition rule leaves a label j satisfying*

$$c_j \geq \alpha \frac{S_c}{n}. \tag{53}$$

Proof. Every observed principal block is invertible. Indeed, $\|M_{II}^{-1}\|_\infty \leq 2$, while $\|(\widetilde{M} - M)_{II}\|_\infty \leq n\varepsilon_0$, so the Neumann-series test applies. The algorithm's fallback is therefore irrelevant in this small-noise regime.

Fix the realized batched path and its final omitted label j. Define

$$\theta_0 = \min\left\{\frac{\eta S_c}{n}, \frac{c_*}{32L}\right\}, \qquad z_0 = \theta_0 \exp(-4L/d). \tag{54}$$

Because $S_c \le nc_*$, the definition of α gives $z_0 \ge \alpha S_c/n$. Suppose for contradiction that $c_j < z_0$.

Fix $\theta \in [z_0, \theta_0]$. By (45), some label outside I_0 has inverse mass at least θ. Consider the last block B containing such a label, let I be the set acquired before that block, and put $V = S \setminus (I \cup B)$. Then

$$j \in V, \qquad c_l < \theta \ (l \in V), \qquad B_0 := \max_{h \in B} s_h \ge \theta. \tag{55}$$

The last inequality follows from $s_h \ge c_h$: in $C_{UU} s_U = c_U$, all off-diagonal entries of C are nonpositive and $C_{hh} \le 1$.

Restrict $CD = I - cm^\top$ to $V \times V$, multiply by $A = C_{VV}^{-1}$, and use the nonpositive cross block. With $\omega = m_V^\top c_V$, this gives

$$(1 - \omega)u \le D_{VV} c_V. \tag{56}$$

Equations (43) and (55) give

$$\omega \le \frac{2L\theta}{c_*} \le \frac{1}{16}, \qquad \|u\|_\infty \le \frac{L\theta}{1-\omega} \le 2L\theta. \tag{57}$$

Together with (49) and $Q\mathbf{1} \le \mathbf{1}$,

$$\max_{g \in U} s_g \le B_0 + 2L\theta \le (1 + 2L)B_0, \qquad \|x\|_1 \le S_c + n(1+2L)B_0. \tag{58}$$

The second inequality is the interpolation scale bound $\|x\|_1 \le S_c + |I| \max_U s$.

Let $E = \widetilde{M} - M$, $\varepsilon = \|E\|_{\max}$, and let $\widetilde{s}_U$ be the noisy residual used by the rule. The relative perturbation identity gives, simultaneously for every acquired set on the data-dependent path,

$$\|\widetilde{s}_U - s_U\|_\infty \le 4\varepsilon \|x\|_1 =: e. \tag{59}$$

Since $B_0 \ge \theta \ge z_0 \ge \alpha S_c/n$, (46) and (58) imply

$$\frac{e}{B_0} \le 4\varepsilon n[\alpha^{-1} + 1 + 2L] \le \frac{d}{4}. \tag{60}$$

Every $h \in B$ had noisy score no larger than the score of the ultimately omitted j. Hence

$$s_h \le s_j + 2e \quad (h \in B), \qquad s_j \ge (1 - d/2)B_0. \tag{61}$$

Apply (51) in (49), use $|B| \le b - 1$, (52), and $s_h \le B_0$:

$$\begin{aligned} s_j &\le |B|\kappa B_0 + u_j \left(1 + \frac{4LB_0}{c_*}\right) \\ &\le (1-d)B_0 + u_j \left(1 + \frac{4LB_0}{c_*}\right). \end{aligned} \tag{62}$$

Combining (61) and (62), and using $\theta \le c_*/(32L)$, yields

$$u_j \ge \frac{(d/2)B_0}{1 + 4LB_0/c_*} \ge \frac{4d}{9}\theta. \tag{63}$$

The j-th coordinate of (56) now gives

$$\sum_{l \in V} D_{jl} c_l \ge \frac{d}{3}\theta. \tag{64}$$

For the fixed final row j, define $F_j(\theta) = \sum_{l: c_l < \theta} D_{jl} c_l$. The set V in (55) is contained in this threshold set, so $F_j(\theta) \ge d\theta/3$ throughout $[z_0, \theta_0]$. Integrating once,

$$\begin{aligned} \frac{d}{3} \log \frac{\theta_0}{z_0} &\le \int_{z_0}^{\theta_0} \frac{F_j(\theta)}{\theta^2}\, d\theta \\ &= \sum_{l: c_l < \theta_0} D_{jl} c_l \left(\frac{1}{\max\{z_0, c_l\}} - \frac{1}{\theta_0}\right) \le \sum_l D_{jl} \le L. \end{aligned} \tag{65}$$

The left side equals $4L/3$, a contradiction. Thus $c_j \ge z_0 \ge \alpha S_c/n$. □

A.4 Decoding and the uniform noise bound

Let $\varepsilon_q = H\delta$. The received critics give

$$\|\widetilde{M} - M\|_{\max} \le \gamma\varepsilon_q, \qquad \|HZ - q_P\|_{\max} \le \varepsilon_q. \tag{66}$$

Suppose first that $\varepsilon_q \le \varepsilon_0$. Proposition 3 applies. Write $J = S \setminus \{j\}$. For every zero-sum row ξ, put $y = \xi M$. Since $yc = 0$,

$$\xi = \sum_{g \in J} y_g \left(C_g - \frac{c_g}{c_j} C_j \right), \qquad \|\xi\|_1 \le K_j \|(\xi M)_J\|_\infty, \tag{67}$$

where

$$K_j = 2\left[(n-2) + \frac{S_c}{c_j} \right] \le \frac{4n}{\alpha}. \tag{68}$$

For a true row $p = P_{sa}$, the simplex fit in (23) has residual at most $(1+\gamma)\varepsilon_q$, because $pM_J = q_J(s,a,\cdot)$. The fitted row $\widehat{p}$ is also stochastic, and a three-term comparison gives

$$\|(\widehat{p} - p)M_J\|_\infty \le 2(1+\gamma)\varepsilon_q \le 4\varepsilon_q. \tag{69}$$

Since $\widehat{p} - p$ sums to zero, (67)–(68) yield

$$\|\widehat{p} - p\|_1 \le \frac{16n}{\alpha} H\delta. \tag{70}$$

If $H\delta > \varepsilon_0$, both rows are probability distributions and their distance is at most two. Using (46), the two regimes combine as

$$\max_{s,a} \|\widehat{P}_{sa} - P_{sa}\|_1 \le \min\left\{ 2, \frac{64}{d\alpha} nH\delta \right\}. \tag{71}$$

This proves (25) for $n \ge b + 2$, with $C_\kappa = 64/(d\alpha)$. Neither the acquisition rule nor the fit uses δ or the thresholds in (46).

It remains to cover $b \ge n - 1$. Query all $n - 1$ retained labels in one round. Here $1 - (n - 2)\kappa > 0$, so the arbitrary-batch inverse-mass lemma gives the sole omitted label a positive fraction $\eta_{\kappa,n-1} S_c$. Applying (67) and the same simplex-fit argument gives $\min\{2, C_\kappa nH\delta\}$ after increasing C_κ if needed. For fixed κ, this increase is uniform because $n \le b + 1$.

Finally, if $n \ge b + 2$, the number of rounds is

$$1 + \left\lceil \frac{n - 1 - b}{b - 1} \right\rceil = \left\lceil \frac{n - 2}{b - 1} \right\rceil. \tag{72}$$

The same expression equals one when $b \ge n - 1$, proving the stated round bound.

A.5 Recovery-to-execution transfer

Let $e = \max_{s,a} \|\widehat{P}_{sa} - P_{sa}\|_1$. Every stationary policy π has true value span

$$\operatorname{span}(W_P^\pi) \le \frac{1}{1 - \gamma\kappa} \le \frac{1}{1 - \kappa}. \tag{73}$$

Probability-row differences cancel constants, so $\|(P_\pi - \widehat{P}_\pi) W_P^\pi\|_\infty \le e/[2(1 - \gamma\kappa)]$. Applying the arrival-value comparison to a true optimal policy and an optimal policy in the fitted model gives

$$\max_o (1-\gamma)\{J_P^*(o) - J_P^{\widehat{\pi}}(o)\} \le \frac{e}{1 - \gamma\kappa} \le \frac{e}{1 - \kappa}. \tag{74}$$

Only the true kernel is required to satisfy the diameter bound. Combining this display with (71) and the automatic loss cap of one proves (31). Taking the better of this planner and the critic-free online guarantee $B_{n,m,\kappa}(H)$, where $m = \sum_s |A_s|$, from Corollary 5 gives the envelope in the main text. Under the common-action convention there, $B_\kappa(H) = B_{n,nA,\kappa}(H)$.

B LOWER BOUNDS FOR QUERY–FEEDBACK DEPTH

This appendix proves the lower bounds in Section 6. We first state the acquisition protocol precisely, then prove a joint lower bound for any query and round budget. The final two subsections treat exact ambiguity below the identifying budget and the complete-library endpoint on every fixed overlap class.

B.1 PROTOCOL AND RISKS

Fix n states, a common set of A action labels, discount $\gamma = 1 - H^{-1}$ with $H \geq 2$, and $0 < \kappa < 1$. An instance has a kernel P satisfying

$$\max_{s,u,x,v} \mathrm{TV}(P_{su}, P_{xv}) \leq \kappa \tag{75}$$

and a complete normalized singleton-avoidance table Z satisfying

$$\|Z - \bar{q}_{P,\mathrm{full}}\|_{\max} \leq \delta, \qquad \bar{q} = q/H. \tag{76}$$

One query returns a whole state–action column of this same table. In each feedback round, the learner chooses a batch of previously unread labels before receiving any answer from that batch. Later batches may depend on all earlier answers. Acquisition observes no transition sample and ends before the downstream reward and initial state are disclosed.

The kernel, the complete table, the downstream reward, and the initial state are fixed before the learner's private random stream. After acquisition, the execution rule may depend on its complete trajectory and may continue learning without restriction. Thus a bound below is not caused by freezing the post-acquisition policy.

Let $\mathfrak{A}_q^{(r)}$ contain randomized acquisition rules using at most q distinct labels in at most r feedback rounds. In the recovery risk, an element also includes its terminal estimator $\widehat{P}$; in the execution risk, it also includes the post-acquisition execution rule. Their recovery risk is

$$\mathcal{R}_q^{(r)}(\delta; H, \kappa) = \inf_{\mathsf{A} \in \mathfrak{A}_q^{(r)}} \sup_{P,Z} \mathbb{E}_{P,Z}^{\mathsf{A}} \max_{s,u} \|\widehat{P}_{su} - P_{su}\|_1, \tag{77}$$

where the supremum is over (75)– (76). Adding a downstream execution rule to the same acquisition class gives

$$\mathcal{D}_q^{(r)}(\delta; H, \kappa) = \inf_{\mathsf{A} \in \mathfrak{A}_q^{(r)}} \sup_{P,Z,\ r_0 \in [0,1]^n,\ o \in S} \frac{1}{H} \left\{ J_P^*(o; r_0) - \mathbb{E}_{P,Z,r_0,o}^{\mathsf{A}} \sum_{k \geq 0} \gamma^k r_0(S_{k+1}) \right\}. \tag{78}$$

The reward is written r_0 to distinguish it from the number of feedback rounds.

It is harmless to assume that an algorithm uses exactly q distinct labels. Embed an early-stopping rule in r round slots. If it reaches the final slot, add unused labels before receiving that slot's replies. If it stops earlier, use one remaining slot for padding and ignore all added replies. This preserves the original output and uses at most r nonempty rounds. We also take an acquisition rule to be defined on arbitrary finite numerical reply histories. A rule initially specified only on feasible histories can be extended arbitrarily elsewhere without changing its behavior on any legal instance. These conventions let us trace the rule on a simple response path below; the eventual lower bound is always evaluated on legal tables.

B.2 THE EXCLUSION-CORE FAMILY

Put

$$b = \left\lceil \frac{1}{\kappa} \right\rceil, \qquad \ell = b + 1, \tag{79}$$

so $b \geq 2$, and assume $n \geq \ell + 1$ and $A \geq \ell + 2$. Choose an ℓ-label core R, a special label $j \in F := S \setminus R$, and a bit $B \in \{0, 1\}$. Write

$$w = \mathrm{Unif}(R), \qquad v_g = \mathrm{Unif}(R \setminus \{g\}) \quad (g \in R), \tag{80}$$

and, for a parameter θ,

$$p_\theta = \theta e_j + (1-\theta)w. \tag{81}$$

At every core state, provide the ℓ exclusion rows v_g and two designated probe rows. Let

$$\theta_0 = \frac{1}{4b^2}, \qquad \theta_\pm = \theta_0 \pm \tau. \tag{82}$$

Under bit zero, the two probe labels have rows $p_{\theta_+}, p_{\theta_-}$; bit one exchanges them. At every state in F, all actions have row w. Additional action labels duplicate an exclusion row on the core and use w on F.

For every value of τ used below, $\tau \le \theta_0/2$. Hence the rows are stochastic and $\theta_+ \le 3/(8b^2) < 1/b^2$. The relevant distances are

$$\begin{aligned} &\mathrm{TV}(v_g, v_h) = \frac{1}{b}, && \mathrm{TV}(v_g, w) = \frac{1}{\ell}, \\ &\mathrm{TV}(v_g, p_\theta) = \frac{1+b\theta}{\ell} \le \frac{1}{b}, && \mathrm{TV}(w, p_\theta) = \theta, \qquad \mathrm{TV}(p_{\theta_+}, p_{\theta_-}) = 2\tau. \end{aligned} \tag{83}$$

Thus every kernel in the family satisfies $\kappa(P) = 1/b \le \kappa$. This calculation covers the whole strict contraction range: for $\kappa > 1/2$, it simply gives $b = 2$ and a three-label core whose diameter is $1/2$.

The minimum-occupation matrix is common to the two bits:

$$M = I + \gamma \sum_{f \in F} e_f w. \tag{84}$$

For a core goal g, action v_g avoids g forever after the current state; a forced state first returns to w. For a forced goal, every future minimum is zero. These observations verify (84) directly from the Bellman equations. On a core probe row, the unnormalized arrival critics are

$$q_j(p_\theta) = \theta, \qquad q_g(p_\theta) = \frac{1 - H^{-1}\theta}{\ell} \quad (g \in R), \qquad q_f(p_\theta) = 0 \quad (f \in F \setminus \{j\}). \tag{85}$$

In particular, the midpoint of every core column is independent of both the bit and the identity of j, while all nonspecial forced columns vanish.

For either of the two calibrations used below, define a complete noisy table by

$$(Z_B)_{-j} = \tfrac{1}{2}(\bar{q}_{0,-j} + \bar{q}_{1,-j}), \qquad (Z_B)_j = \bar{q}_{B,j}. \tag{86}$$

The special column is exact. Each core column has maximum entrywise error $\tau/(\ell H^2)$, while every nonspecial forced-goal column is exact. Hence the maximum error over all columns other than j is

$$\frac{\tau}{\ell H^2}. \tag{87}$$

Thus (86) is legal whenever $\tau \le \ell \delta H^2$.

B.3 A JOINT LOWER BOUND FOR QUERIES AND ROUNDS

For $0 \le q \le n$ and $r \ge 1$, define

$$\Psi_{q,r,\ell} = \min_{\substack{k_1+\cdots+k_r=q \\ k_t \in \mathbb{N}_0}} \sum_{t=1}^{r} \binom{k_t}{\ell} \tag{88}$$

and

$$\pi_{n,q,r,\ell} = \frac{n-q}{n-\ell} \frac{\Psi_{q,r,\ell}}{\binom{n}{\ell}}. \tag{89}$$

We use the convention $\binom{k}{\ell} = 0$ for $k < \ell$; at $q = n$, the factor $n - q$ makes $\pi_{n,q,r,\ell} = 0$.

Proposition 4 (Joint exhaustion lower bound)**.** *Under the conditions of Section B.2, for every $0 \le q \le n$, $r \ge 1$, and $\delta \ge 0$,*

$$\mathcal{D}_q^{(r)}(\delta; H, \kappa) \ge \frac{\pi_{n,q,r,\ell}}{5} \min\left\{ \ell\delta H^2, \frac{1}{32b^2\sqrt{H}} \right\}, \tag{90}$$

$$\mathcal{R}_q^{(r)}(\delta; H, \kappa) \ge 2\pi_{n,q,r,\ell} \min\left\{ \ell\delta H^2, \frac{1}{8b^2} \right\}. \tag{91}$$

Both statements hold for randomized batch sizes and adaptive stopping.

Proof. Fix an algorithm, pad it to exactly q queries, and let U contain its entire private stream. Draw R uniformly among the ℓ-subsets of S, then draw j uniformly from $S \setminus R$ and an independent fair bit B. These variables determine the complete kernel and table before U is drawn. For the execution statement, also set the reward to e_j and the initial state to the least element of R under a fixed ordering.

For a fixed stream u, run acquisition on a response path that returns a zero column at every query. Let its disjoint batches be $B_1^0(u), \dots, B_r^0(u)$, with sizes $k_1(u), \dots, k_r(u)$ summing to q. Define

$$\mathcal{G}(u) = \{R : |R| = \ell,\ R \subseteq B_t^0(u) \text{ for some } t\}. \tag{92}$$

The batches are disjoint, so each such core lies in a unique batch and

$$|\mathcal{G}(u)| = \sum_{t=1}^{r} \binom{k_t(u)}{\ell} \geq \Psi_{q,r,\ell}. \tag{93}$$

For each fixed core R, trace the same algorithm on replies that return the common midpoint on core columns and zero outside R. Let $J_R(u)$ be the final queried set. This response path depends on R, but not on j or B, and $|J_R(u)| = q$. If $R \in \mathcal{G}(u)$, all batches before the unique core-containing batch avoid R and therefore receive zeros on both paths. The two paths agree until that entire batch has been committed. Consequently,

$$R \subseteq J_R(u). \tag{94}$$

Let

$$E = \{R \in \mathcal{G}(U),\ j \notin J_R(U)\}. \tag{95}$$

For an exhausted core, $J_R(u)$ contains ℓ core labels and $q - \ell$ outside labels. Since j is uniform among the $n - \ell$ outside labels and $J_R(u)$ is independent of j,

$$\Pr(E \mid U = u) = \frac{|\mathcal{G}(u)|}{\binom{n}{\ell}} \frac{n-q}{n-\ell} \geq \pi_{n,q,r,\ell}. \tag{96}$$

On E, the actual legal table (86) and the core-midpoint response path agree at every queried label. This follows inductively across batches: the only differing column is j, and that label is never queried. The auxiliary response path need not itself be a legal complete table; every observed response on the event is generated by the legal instance.

Fix R, j for which the event has positive probability. The event depends on R, j, U, but not on B. Conditional on R, j, E, the bit remains fair, the conditional law of the entire private stream is common to the two bits, and the acquired labels and replies agree. Acquisition observed no transition, so this conditioning does not alter either model's future transition law.

For recovery, choose

$$\tau_{\mathrm{rec}} = \min\left\{\ell\delta H^2, \frac{\theta_0}{2}\right\} = \min\left\{\ell\delta H^2, \frac{1}{8b^2}\right\}. \tag{97}$$

The two kernels exchange the two probe rows on every core state. For either probe label,

$$\|p_{\theta_+} - p_{\theta_-}\|_1 = \|2\tau_{\mathrm{rec}}(e_j - w)\|_1 = 4\tau_{\mathrm{rec}}. \tag{98}$$

The estimator has the same conditional distribution under both bits. For every realized output, the triangle inequality makes the average of its two maximum-row errors at least $2\tau_{\mathrm{rec}}$. Multiplying by the miss probability (96), averaging the fixed finite prior, and choosing a fixed instance at least as hard as that average proves (91).

For execution, choose

$$\tau_{\mathrm{exec}} = \min\left\{\ell\delta H^2, \frac{\theta_0}{8\sqrt{H}}\right\} = \min\left\{\ell\delta H^2, \frac{1}{32b^2\sqrt{H}}\right\}. \tag{99}$$

For either direction of comparison, one probe transition has divergence

$$\mathrm{KL}(p_{\theta_+} \| p_{\theta_-}) \leq \frac{4\tau_{\mathrm{exec}}^2}{\theta_-(1-\theta_-)} \leq \frac{8\tau_{\mathrm{exec}}^2}{\theta_0}, \tag{100}$$

and the reverse divergence obeys the same bound. All other rows agree between bits. Through $T = \lceil H \rceil \le 2H$ decisions, the adaptive KL chain rule gives history divergence at most

$$\frac{16H\tau_{\mathrm{exec}}^2}{\theta_0} \le \frac{\theta_0}{4} \le \frac{1}{64}. \tag{101}$$

Pinsker's inequality bounds the corresponding total variation by $1/8$. This calculation begins from the common conditional private-stream law and therefore remains valid after conditioning on E.

For reward e_j, the larger probe is optimal at every core state. Its arrival advantage over the smaller probe is

$$\frac{2\tau_{\mathrm{exec}}}{1+\gamma\theta_+} \ge \frac{8\tau_{\mathrm{exec}}}{5}, \tag{102}$$

and every exclusion action is at least as suboptimal. At each decision the state belongs to the core with probability at least $3/4$. Because the two bits exchange the optimal probe label, the testing inequality and (101) give bit-averaged probability at least $(3/4 - 1/8)/2 \ge 1/4$ of choosing a suboptimal core action. Finally,

$$H^{-1}\sum_{k=0}^{T-1}\gamma^k \ge \frac{1}{2}. \tag{103}$$

The arrival performance identity therefore gives conditional normalized loss at least

$$\frac{8\tau_{\mathrm{exec}}}{5}\cdot\frac{1}{4}\cdot\frac{1}{2} = \frac{\tau_{\mathrm{exec}}}{5}.$$

This argument permits every action to depend on the entire preceding trajectory. Multiplying by (96), averaging the fixed prior, and selecting a fixed instance proves (90). □

The coefficient has a closed form. Write

$$a = \lfloor q/r \rfloor, \qquad s = q - ra, \qquad 0 \le s < r. \tag{104}$$

Since the discrete increment of $k \mapsto \binom{k}{\ell}$ is $\binom{k}{\ell-1}$, moving one query from a batch that is larger by at least two to the smaller batch cannot increase the sum. Repeating this balancing step gives

$$\Psi_{q,r,\ell} = (r-s)\binom{a}{\ell} + s\binom{a+1}{\ell}. \tag{105}$$

It follows that

$$\pi_{n,q,r,\ell} > 0 \quad\Longleftrightarrow\quad q < n \ \text{ and } \ q > r(\ell-1) = rb. \tag{106}$$

For $q = n-1$, (89) reduces to $p_{n,r,\ell}$ in (33), proving Theorem 4. At $r = 1$ it equals $1/n$.

B.4 Exact ambiguity below the identifying budget

The exclusion-core bound is designed for noisy stability. With two or more unread labels, a simpler family gives exact ambiguity. Assume $A \ge 2$. Choose a hidden unordered pair $\{i, j\}$ uniformly, let $u = \mathbf{1}/n$, and define

$$p_+ = u + t(e_i - e_j), \qquad p_- = u - t(e_i - e_j). \tag{107}$$

At every state, model zero assigns p_+ and p_- to two action labels; model one exchanges those labels. Extra labels duplicate the first action in each model. The hidden pair and an independent fair bit are drawn before the learner's private stream.

For any goal g, put $\mu_g = \min_v p_v(g)$. Because the rows do not depend on the current state, Bellman uniqueness gives

$$q_g(s,v) = p_v(g) + \gamma H\mu_g, \qquad \bar{q}_g(s,v) = H^{-1}p_v(g) + \gamma\mu_g. \tag{108}$$

For every $g \notin \{i, j\}$, the entire exact normalized column is the constant $1/n$, independently of the pair and bit.

Fix $0 \le q \le n-2$ and an algorithm using at most q labels in any number of rounds. For each private stream, trace it on the constant-column response path and let J be the resulting queried set. If the hidden pair is disjoint from J, the actual exact responses follow the same path. Therefore

$$\Pr(\{i,j\} \cap J = \varnothing \mid U) = \frac{\binom{n-|J|}{2}}{\binom{n}{2}} \ge \frac{\binom{n-q}{2}}{\binom{n}{2}}. \tag{109}$$

This event depends on the pair and private stream, but not on the bit, so the conditional private-stream law is common to both models.

For recovery, take

$$t_{\mathrm{rec}} = \min\left\{\frac{1}{4n}, \frac{\kappa}{4}\right\}. \tag{110}$$

The rows are stochastic and each model has diameter $2t_{\mathrm{rec}} \le \kappa$. The two labeled kernels are separated by $4t_{\mathrm{rec}}$ in maximum-row ℓ_1 distance. The conditional output law is common, so the triangle inequality and (109) give

$$\mathcal{R}_q^{(r)}(0; H, \kappa) \ge 2t_{\mathrm{rec}} \frac{\binom{n-q}{2}}{\binom{n}{2}} > 0. \tag{111}$$

For execution, use reward e_i, any fixed initial state, and

$$t_{\mathrm{exec}} = \min\left\{\frac{1}{16n\sqrt{H}}, \frac{1}{4n}, \frac{\kappa}{4}\right\}. \tag{112}$$

For either direction, the one-step divergence is at most $8nt_{\mathrm{exec}}^2$. Through $\lceil H \rceil$ decisions, history divergence is at most $16nHt_{\mathrm{exec}}^2 \le 1/(16n)$, and total variation is at most $1/8$. The two bits have disjoint sets of optimal action labels; a suboptimal decision loses exactly $2t_{\mathrm{exec}}$ in expected arrival reward. The same testing and discounted-sum calculation as above yields conditional normalized loss at least $t_{\mathrm{exec}}/4$. Thus

$$\mathcal{D}_q^{(r)}(0; H, \kappa) \ge \frac{t_{\mathrm{exec}}}{4} \frac{\binom{n-q}{2}}{\binom{n}{2}} > 0. \tag{113}$$

Equations (111) and (113) prove Proposition 2. The number of acquisition rounds does not appear: feedback cannot reveal a column that is never queried.

B.5 The complete-library benchmark on a fixed overlap class

We close the comparison by proving that the complete library has the stated minimax horizon–noise order on every fixed strict-diameter class.

Proposition 5 (Full-library recovery lower bound)**.** *For every $n \ge 3$, $A \ge 1$, $0 < \kappa < 1$, $H \ge 2$, and $\delta \ge 0$,*

$$\mathcal{R}_n^{(1)}(\delta; H, \kappa) \ge \frac{\kappa}{2} \min\{1, nH\delta\}. \tag{114}$$

Proof. Reserve a root o. Choose an even set $E \subseteq S \setminus \{o\}$ of size

$$N = 2\left\lfloor \frac{n-1}{2} \right\rfloor, \tag{115}$$

an anchor $z \in E$, and balanced signs $\sigma_i \in \{-1, 1\}$ on E. For $0 \le t \le 1$, let

$$p_\pm(i) = \frac{1 \pm t\sigma_i}{N} \quad (i \in E) \tag{116}$$

and zero elsewhere. Every nonroot row is e_z, while the root row is

$$r_\pm = (1-\kappa)e_z + \kappa p_\pm. \tag{117}$$

One action suffices; duplicate it for any larger prescribed action count. Within either model, $\mathrm{TV}(r_\pm, e_z) = \kappa\,\mathrm{TV}(p_\pm, e_z) \le \kappa$.

No transition returns to o. For singleton goal g,

$$V_g(z) = H\mathbf{1}\{g = z\}, \qquad V_g(s) = \mathbf{1}\{s = g\} + \gamma H\mathbf{1}\{g = z\} \quad (s \neq o). \tag{118}$$

All nonroot critic rows therefore agree between models, while

$$q_g^{\pm}(o) = r_{\pm}(g) + \gamma H\mathbf{1}\{g = z\}. \tag{119}$$

It follows that

$$\|\bar{q}_+ - \bar{q}_-\|_{\max} = \frac{2\kappa t}{NH}, \qquad \|r_+ - r_-\|_1 = 2\kappa t. \tag{120}$$

Choose

$$t = \min\left\{1, \frac{NH\delta}{\kappa}\right\}, \qquad Z = \tfrac{1}{2}(\bar{q}_+ + \bar{q}_-). \tag{121}$$

The common input Z is within δ of both complete tables. Coupling decoder randomness and applying the triangle inequality to the root row forces error at least $\kappa t = \min\{\kappa, NH\delta\}$. Since $N \geq n/2$ and $\min\{\kappa, x\} \geq \kappa \min\{1, x\}$, this is at least the right side of (114). □

The same exclusion-core family gives the corresponding execution endpoint.

Proposition 6 (Full-library execution lower bound)**.** *Fix $0 < \kappa < 1$ and let $b = \lceil 1/\kappa \rceil$. If $n \geq b + 2$, $A \geq b + 3$, $H \geq 2$, and $\delta \geq 0$, then*

$$\mathcal{D}_n^{(1)}(\delta; H, \kappa) \geq \frac{1}{160b^2} \min\{\delta H, H^{-1/2}\}. \tag{122}$$

Proof. Fix one core R and one special label j in the family of Section B.2. This time give the learner the midpoint of the two *complete* normalized tables. By (85), their maximum distance is $2\tau/H$, attained in the special column. Set

$$\tau = \min\left\{\delta H, \frac{\theta_0}{8\sqrt{H}}\right\}. \tag{123}$$

The midpoint is legal under either bit, so the acquisition transcript and private-stream law agree. The history comparison (100)–(103) applies without conditioning and gives bit-averaged loss at least $\tau/5$. Therefore

$$\mathcal{D}_n^{(1)}(\delta; H, \kappa) \geq \min\left\{\frac{\delta H}{5}, \frac{\theta_0}{40\sqrt{H}}\right\}. \tag{124}$$

Since $\theta_0 = 1/(4b^2)$, the smaller coefficient is $1/(160b^2)$, proving (122). □

Together with Corollaries 7, 8, and 5, Propositions 5 and 6 show that the benchmark used by the feedback-depth theorem is the minimax horizon–noise order on the same fixed-κ class. The last critic therefore changes the required acquisition depth while leaving the optimal recovery order unchanged.

C Finite recovery checks

This appendix proves the observed-data certificate used in Section 7, describes the prespecified four-action study, and reports a complementary check on the lower-bound construction. The numerical results test the implementations and illustrate finite-instance behavior; the minimax laws are proved analytically.

C.1 Observed-data recovery certificate

The uniform constant in Theorem 3 is conservative. An acquired set and its fitting residual give a separate bound for the particular input.

Corollary 4 (Observed recovery bound). *Let I contain $n-1$ distinct acquired labels, let $\widehat{P}$ have probability rows, and suppose* (3) *holds. Use $\widetilde{M}_g = e_g + \gamma H \min_u Z_g(\cdot, u)$ and define*

$$\mathsf{T}_I = [\widetilde{M}_I, \mathbf{1}], \qquad \widehat{\rho}_I = \max_{s,u} \|\widehat{P}_{su}\widetilde{M}_I - HZ_I(s,u,\cdot)\|_\infty. \tag{125}$$

If T_I is invertible, put

$$\widehat{K}_I = \sum_{k=1}^{n-1} \|(\mathsf{T}_I^{-1})_{k,:}\|_1, \qquad \mathcal{E}_I = \min\{2, \widehat{K}_I[\widehat{\rho}_I + (1+\gamma)H\delta]\}.$$

Otherwise set $\mathcal{E}_I = 2$. Then

$$\max_{s,u} \|\widehat{P}_{su} - P_{su}\|_1 \le \mathcal{E}_I. \tag{126}$$

The set I and fit may use the same noisy table. No transition-diameter condition or optimality of the fit is required.

Proof. The Bellman identity and the Lipschitz property of the minimum give $\|\widetilde{M}_I - M_I\|_{\max} \le \gamma H\delta$ and

$$\|P_{su}\widetilde{M}_I - HZ_I(s,u,\cdot)\|_\infty \le (1+\gamma)H\delta.$$

Thus, for $v = \widehat{P}_{su} - P_{su}$, $\|v\widetilde{M}_I\|_\infty \le \widehat{\rho}_I + (1+\gamma)H\delta$. Since $v\mathbf{1} = 0$, invertibility implies $v = [v\widetilde{M}_I, 0]\mathsf{T}_I^{-1}$. The triangle inequality gives the stated coefficient; probability rows give the cap 2. These identities hold for every admissible table and every acquired set, including a set chosen from that table. □

The certificate is conditional on the declared joint error budget. It uses the acquired columns and the output kernel; it neither reads the omitted column nor estimates δ from the fitting residual. A large value of $\mathcal{E}_I$ can be uninformative without implying large actual error.

C.2 Prespecified four-action evaluation

The protocol was fixed before the evaluation seed was drawn. It uses $n = 8$, $A = 4$, $\kappa \in \{1/3, 1/4\}$, and $H \in \{16, 64, 256\}$. For each family and diameter, 16 independently generated kernels form the statistical units. Horizons, noise shapes, and amplitudes are paired repeated measurements within a kernel.

The *weighted punctured* family gives every state–action row a row-dependent missing next-state coordinate. It interpolates between the uniform distribution on the remaining coordinates and a lognormally weighted, independently jittered distribution on the same support. The interpolation weight is found by bisection so that the global row diameter equals κ. The missing-coordinate schedule covers every next-state label, so each coordinate is zero in some row and the family has zero common mass. It has no hidden core, probe action, or distinguished diagnostic state. The *local-with-teleportation* family mixes a random common distribution with a four-action local kernel on a directed cycle, again scaled to the specified diameter. It is a structured common-mass control. A fixed random permutation is applied to states and critic labels before acquisition.

For each kernel and horizon, the study includes one table with no added noise and two noise shapes at normalized amplitudes $\epsilon \in \{2^{-10}, 2^{-8}, 2^{-6}\}$. The low-rank direction is

$$E(s,a,g) = \frac{\epsilon}{H}\big(0.55G_g + 0.25S_s + 0.20A_a\big),$$

with independent coordinates in $[-1,1]$; the second direction is an entrywise Rademacher tensor scaled by ϵ/H. Each complete noisy table is fixed before acquisition and shared by every method. The primary range is $0 < \epsilon \le 2^{-8}$; 2^{-6} is a mandatory saturation check. The declared joint budget is

$$\delta = \frac{\epsilon}{H} + \delta_{\text{generation}}, \qquad \delta_{\text{generation}} = \max\{r_{\text{PI}}, r_{\text{joint}}\} + \frac{128\epsilon_{\text{mach}}n}{1-\gamma},$$

where r_{PI} is the largest per-goal policy-iteration Bellman residual, r_{joint} is the final joint Bellman residual, and ϵ_{mach} is double-precision machine epsilon. Thus entries with $\epsilon = 0$ are zero-added-noise cases, not zero-budget cases. This gives 1,344 conditions in all and 768 primary conditions.

The block and one-column rules share the initial b labels, residual score, tie rule, seven-column budget, and row-wise fit. Only the response schedule differs. For Figure 1(a), each family–diameter–horizon–kernel value first takes the median normalized error over the two noise shapes and two primary amplitudes; the median and 10th–90th percentiles are then taken across the 16 kernels. The recovery endpoint requires a block-to-one-column error ratio at most two in at least 90% of each passing kernel's primary conditions, with at least 14 of 16 kernels passing. Every horizon–noise–amplitude stratum must have at least 75% such hits and median ratio at most two. Ratios add 10^{-8} to both errors before division. Certificate criteria are applied separately within each family–diameter–horizon–noise-shape stratum. For each kernel, its two primary amplitudes first give the fractions with $\mathcal{E}_I < 2$ and $\mathcal{E}_I \le 0.5$, and the median $\mathcal{E}_I/e$ over errors $e > 10^{-8}$, when defined. Across the 16 kernels, the two fractions are averaged and the defined tightness summaries are reduced to their median and 90th percentile. Each stratum requires these averaged fractions to be at least 0.90 and 0.75, at least eight defined tightness summaries, and tightness median and 90th percentile at most 12 and 20. All criteria and aggregation rules were fixed before data generation.

Table 2: Primary four-action results. Each row contains 16 independent kernels and 192 paired low-noise conditions. "Hits" counts conditions with stabilized block-to-one-column error ratio at most two; "Kernels" counts kernels with such hits in at least 90% of their conditions. "Median ratio" is the median of the 16 kernel medians. The last column is the largest kernel-level 90th-percentile certificate/error ratio among the six horizon–noise strata.

Family	κ	Rounds	Hits	Kernels	Median ratio	$\mathcal{E}_I \le .5$	Max q90
Local control	$1/3$	$3/5$	$192/192$	$16/16$	1.000	$192/192$	7.28
Local control	$1/4$	$2/4$	$192/192$	$16/16$	1.000	$192/192$	5.39
Punctured stress	$1/3$	$3/5$	$192/192$	$16/16$	1.000	$192/192$	11.74
Punctured stress	$1/4$	$2/4$	$186/192$	$15/16$	1.000	$192/192$	12.27

All twelve recovery strata in each row met the stratum criteria above. All certificate strata met their criteria as well; their median certificate-to-error ratios ranged from 1.29 to 9.83. At the saturation amplitude, every block fit remained within a factor two of its paired one-column fit. The methods still need not agree pointwise: across the primary range, one punctured-family case had ratio 2.44, and the prespecified recovery criterion allows such finite exceptions.

Control performance was less uniform. A low-noise case entered the control comparison when the median of the eight leave-one-out maximum losses over the fixed reward suite was at least $\max\{10^{-8}, 0.01nH\delta\}$. The suite includes a near-switch panel constructed from the reference kernel only for evaluation. For each kernel, we computed the fraction of included conditions for which the block loss was at most twice the paired one-column loss plus 10^{-8}, and then averaged these fractions equally across kernels; the threshold was 0.90. This prespecified control-loss criterion was met in both local-control rows and in the punctured $\kappa = 1/4$ row. The punctured $\kappa = 1/3$ row reached 0.896. The punctured family also did not consistently improve over omissions fixed before the noisy table was seen. Thus the study supports the recovery and certificate claims in the main text, but not a general control or omission-superiority claim.

C.3 Construction-specific mechanism check

We also report a finite check on the hidden-core family used in the lower bound. It fixes $n = 8$, $A = 7$, $\kappa \in \{1/3, 1/4\}$, and $H \in \{16, 256, 4096\}$, and varies two model bits, three label layouts, and six amplitudes of a common correlated perturbation. This gives 216 conditions. The aligned layout exposes the entire core in a fixed $(b+1)$-label prefix and leaves the special label for last. Two controls delay a core label or reveal the special label in the first batch.

The construction fixes $\theta_0 = 1/(4b^2)$ and $\tau = \theta_0/4 = 1/(16b^2)$.

In every condition, block and one-column feedback retained the same seven labels and produced the same fitted kernel, recovery error, and certificate. Thus only the response schedule changed: $[3, 2, 2]$ versus $[3, 1, 1, 1, 1]$, or $[4, 3]$ versus $[4, 1, 1, 1]$. In the aligned conditions without the additional correlated perturbation, a fixed comparison protocol whose first batch contains $b + 1$ labels omits

the special label, whereas the block rule omits a core label. Direct calculation gives

$$e_{\text{block}} = \frac{2b\tau}{(b+1)H}, \qquad e_{b+1} = 2\tau, \qquad \frac{e_{b+1}}{e_{\text{block}}} = \frac{(b+1)H}{b}. \tag{127}$$

The two control layouts do not force the special omission. This check isolates the mechanism on the lower-bound family; the four-action study above supplies the out-of-family comparison.

An independent implementation regenerated all experimental quantities and reproduced every reported result across 1,344 conditions and 18,816 method records. A separate implementation recomputed 48,384 row-level linear programs for the hidden-core check. The supplement provides the protocols, code, complete results, and replication scripts.

D CHOOSING THE NEXT QUERY USING RECEIVED CRITICS

For each queried goal, compute its noisy occupation column $\widetilde{M}_g = e_g + \gamma H \min_a Z_g(\cdot, a)$. Let I contain the queried labels and $U = S \setminus I$. Interpolate the constant vector at states I using the acquired columns:

$$\widehat{x} = (\widetilde{M}_{II})^{-1}\mathbf{1}, \qquad \widetilde{b}_U = \mathbf{1} - \widetilde{M}_{UI}\widehat{x}. \tag{128}$$

The next query selects an unseen label with the smallest signed residual. Although its index lies in U, every matrix entry used in this score belongs to a column in I.

Sequential critic acquisition

1. Acquire the prescribed initial set I_0 and set $I = I_0$. If $|I| = n - 1$, proceed directly to fitting.
2. While $|I| < n-1$, compute the scores in (128), using scores 1 when I is empty. Query an unqueried label of minimum signed score, break ties by a fixed order, and add its column to I. If the required principal block is singular, query the first unqueried label in that order instead.

Return the queried columns and their labels. The initial sets covered by the guarantees are specified in Theorems 5 and 6. Selection uses neither the downstream reward nor the error budget.

Fit each transition row by minimizing the maximum absolute Bellman residual over the probability simplex:

$$\widehat{P}_{su} \in \arg\min_{p \in \Delta(S)} \|p\widetilde{M}_I - HZ_I(s,u,\cdot)\|_\infty. \tag{129}$$

This fit also needs no error budget. The guarantee applies to the same noisy table that determined which columns were acquired.

How an observed column guides the next query. Partition the states into a three-state core R and states F that return to its uniform row w. The construction in Appendix I has

$$M = \begin{pmatrix} I_3 & 0 \\ \gamma \mathbf{1}_{|F|} w & I_{|F|} \end{pmatrix}, \qquad c = \begin{pmatrix} \mathbf{1}_3 \\ H^{-1}\mathbf{1}_{|F|} \end{pmatrix}. \tag{130}$$

With exact input and $\gamma > 0$, a first core query leaves residual 1 at the other core states and $1 - \gamma/3$ at each state in F. The rule therefore queries F before consuming the remaining core labels. It leaves a core label with $c_j = 1$ unqueried. Theorem 5 extends this protection to every kernel in the contracting class, with the complete library's recovery order under noise.

The proof gives the absolute constants

$$L_0 = 32, \qquad \alpha_0 = 2^{-266}, \qquad C_{\text{rec}} = 2^{273}. \tag{131}$$

These constants establish the optimal parameter dependence; the resulting numerical bound is very conservative.

A shared occupation budget. Put $m_g = \min_s M_{sg}$ and $D = M - \mathbf{1}m^\top$. If $\kappa(P) \le \kappa < 1$, each column satisfies $D_{gg} \ge (1+\gamma\kappa)^{-1} \ge 1/2$ and $0 \le D_{sg} \le \gamma\kappa D_{gg}$ for $s \ne g$. Separate optimization also imposes a constraint on their sum.

Lemma 1 (Joint centered occupation)**.** *Suppose $0 < \kappa < 1$ and $\kappa(P) \le \kappa$. Define*

$$K_\kappa = \min\left\{k \ge 1 : k\kappa^{k-1} \le \frac{1-\kappa}{4}\right\},$$
$$\beta_\kappa = \frac{3+\kappa}{4}, \qquad \eta_\kappa = \frac{3(1-\kappa)}{4}, \tag{132}$$
$$L_\kappa = \frac{K_\kappa}{\beta_\kappa}\left(1 + \frac{1}{\eta_\kappa}\right).$$

Then

$$\max_s \sum_g D_{sg} \le L_\kappa < \infty. \tag{133}$$

If $\kappa \le 1/2$, the six-step specialization gives $\max_s \sum_g D_{sg} \le 2016/65 < L_0$.

For each goal, start an optimal trajectory at a state minimizing that goal's value, and couple it to one common reference trajectory. Until they meet, the excess cost is bounded by the reference's visits to that goal. The reference periodically regenerates from a common distribution; each regeneration gives every unmatched goal a fixed chance to meet it. However, the goals that remain unmatched can depend on the new reference state, so multiplying their survival probabilities by its occupation probabilities is invalid.

We charge the next interval to the goals that were unmatched *before the regeneration block began*. This older set is independent of the new reference state and has already passed through all previous matching opportunities. The reference occupies one state at a time. Summing over its locations, then over regeneration intervals, therefore gives a constant total cost even though the goals use different optimal policies.

Protecting an omission across all scales. Write $S_c = \sum_g c_g$ and $c_* = \max_g c_g$. If the initial m goals' minimizing rows share at least $\rho = 1-(m-1)\kappa > 0$ mass, their centered Bellman equations keep the fraction

$$\eta^+_{\kappa,m} = \min\left\{\frac{1}{2m+1}, \frac{\rho(1-\kappa)}{6(m+1)}\right\} > 0 \tag{134}$$

of S_c outside that batch (Lemma 3). Under the diameter-$1/2$ assumption of Theorem 5, any initial pair I_0 has common mass at least $1/2$, and the sharper bound

$$\sum_{g \notin I_0} c_g \ge S_c/16. \tag{135}$$

This prevents both queried labels from carrying all substantial inverse mass. Together with the joint occupation budget, it lets the query rule protect a fraction of the average.

Lemma 2 (Stable omission under noise)**.** *For the rule in Theorem 5, if $\|\widetilde{M} - M\|_{\max} \le \alpha_0/(64n)$, no singular-block fallback occurs and the final omitted label satisfies $c_j \ge \alpha_0 S_c/n$.*

To see the global comparison, set $\theta_0 = \min\{S_c/(16n), c_*/(32L_0)\}$ and $z_0 = \theta_0 2^{-8L_0} \ge \alpha_0 S_c/n$. Suppose the final omitted label has $c_j < z_0$. For each level $\theta \in [z_0, \theta_0]$, consider the last query whose inverse mass is at least θ. Such a query exists by (135), and all labels remaining afterward have smaller mass. The inverse signs and the residual comparison imply

$$F_j(\theta) := \sum_{g: c_g < \theta} D_{jg} c_g \ge \frac{\theta}{5}. \tag{136}$$

Integrating over the levels gives

$$\frac{1}{5}\log\frac{\theta_0}{z_0} \le \int_{z_0}^{\theta_0} \frac{F_j(\theta)}{\theta^2}\, d\theta \le \sum_g D_{jg} \le L_0. \tag{137}$$

The left side exceeds L_0, a contradiction. Each entry contributes at most D_{jg} to the integral, regardless of the number of queries. The constant occupation budget therefore controls the full query sequence.

Selection from a noisy table. For an acquired set I, exact interpolation gives $x = M_{II}^{-1}\mathbf{1}$, $b_U = C_{UU}^{-1}c_U$ and $X = M_{UI}M_{II}^{-1} \geq 0$ with $X\mathbf{1} \leq \mathbf{1}$. Writing $E = \widetilde{M} - M$ and $\varepsilon = \|E\|_{\max}$,

$$\widetilde{b}_U - b_U = (XE_{II} - E_{UI})\widehat{x}, \qquad \|\widetilde{b}_U - b_U\|_\infty \leq 4\varepsilon\|x\|_1 \tag{138}$$

when $2|I|\varepsilon \leq 1/2$. The bound uses $\|\widehat{x}\|_1 \leq 2\|x\|_1$ and retains total mass: $\|x\|_1 \leq S_c + |I| \max_U b$. If g is the last queried label with $c_g \geq \theta$, every label left afterward has inverse mass below θ. This implies $\max_U b \leq b_g + 2L_0\theta$ and $b_g \geq \theta$. The prescribed noise threshold then keeps $b_j \geq 3b_g/4$ at that actual query, which yields (136). The cancellation removes large common values before comparing scores. All inequalities hold for every acquired set, so reuse of the same correlated table throughout acquisition is allowed. Appendix E proves both lemmas and the decoding bound, then extends that bound to all error levels using the fact that the ℓ_1 distance between probability rows is at most 2. For a general permitted initial batch, the same argument replaces $1/16$ by $\eta^+_{\kappa,m}$; this gives the recovery part of Theorem 6.

From recovery to execution. Under the true kernel, every stationary policy's current-state value has span at most $(1-\gamma/2)^{-1} \leq 2$. Planning optimally for arrival rewards in a fitted stochastic model with maximum row error e gives a stationary policy with normalized loss at most $2e$. Only the true model needs to contract. Choosing between this planner and the online learner using the declared δ gives (190). This transfers the linear recovery bound into the critic branch of the execution upper bound; the execution rate comparison in Theorem 6 fixes n, A, κ.

E Linear stability of sequential acquisition

We prove Lemmas 1 and 2, and Theorem 5, then prove the matching recovery lower bound. The parameterized occupation and prefix-mass lemmas hold for $n \geq 2$ and $0 \leq \gamma < 1$. The stable sequential-omission lemma, the diameter-$1/2$ sequential theorem, and its matching lower bound assume $n \geq 4$. Unless a statement carries its own parameter κ, the total-variation diameter of all legal state–action transition rows is at most $1/2$. Action sets are finite and nonempty. Write $a = 1 - \gamma$ and $H = a^{-1}$. The first two queried labels may be prescribed, or may arise from the rule started with an empty set. The proof applies to either case: the initialization bound holds for every two-label set.

E.1 Inverse structure and centered columns

Let M be the true singleton minimum-occupation matrix. Theorem 2 and Appendix J give

$$C = M^{-1}, \qquad c = C\mathbf{1}, \qquad 0 < C_{ii} \leq 1, \quad C_{ij} \leq 0\ (i \neq j), \quad a\mathbf{1} \leq c \leq \mathbf{1}, \quad \|C\|_\infty \leq 2. \tag{139}$$

Matrix infinity norms are maximum absolute row sums. Set

$$m_g = \min_s M_{sg}, \qquad D = M - \mathbf{1}m^\top, \qquad \Delta_g = D_{gg}, \qquad c_* = \max_s c_s, \qquad S_c = \sum_s c_s. \tag{140}$$

Column g attains its maximum only at g. Its maximum is positive because the immediate cost at g is one; a non-goal state attaining that maximum would have value at most γ times itself. Thus a minimum is attained at some $s_g \neq g$.

More generally, suppose the row diameter is at most $\kappa < 1$. The expectations of any vector under two legal rows then differ by at most κ times its span. The same bound holds between the minimum expectations at two states. Subtracting the Bellman equation at s_g from the equations at other states gives

$$0 \leq D_{sg} \leq \gamma\kappa\Delta_g \quad (s \neq g), \qquad 1 - \gamma\kappa\Delta_g \leq \Delta_g \leq 1 + \gamma\kappa\Delta_g. \tag{141}$$

Consequently,

$$\frac{1}{2} \leq \frac{1}{1+\gamma\kappa} \leq \Delta_g \leq \frac{1}{1-\gamma\kappa}. \tag{142}$$

Centering $CM = I$ yields

$$CD = I - cm^\top. \tag{143}$$

We also need the nonnegative row $d^\top = m^\top C$. For each g, choose a row r attaining $M_{rg} = m_g$. Then

$$d_g - (M_{r,:}C)_g = \sum_{h \neq g} (m_h - M_{rh}) C_{hg} \geq 0,$$

so $d_g \geq 0$. Define

$$t_f = 1 - m^\top c = 1 - d^\top \mathbf{1}. \tag{144}$$

Since $Mc = \mathbf{1}$,

$$Dc = t_f \mathbf{1}, \qquad t_f \geq \Delta_g c_g > 0, \qquad d^\top \mathbf{1} < 1. \tag{145}$$

These identities retain the column minima, which need not be small.

E.2 A DIMENSION-FREE JOINT OCCUPATION BUDGET

We use a common source trajectory to compare values optimized under separate policies. A block regeneration refreshes this source from one fixed distribution. We charge the following cycle to the goals that were unmatched *before the regenerating block began*; this delay makes source occupation independent of the charged set. Splitting a minorized kernel is a classical regeneration construction (Nummelin, 1978; Athreya & Ney, 1978). We give the common-mass estimate and the joint accounting argument explicitly.

Proof of Lemma 1. Fix a deterministic stationary source policy with transition matrix P_0, and a stationary distribution μ of P_0. The row diameter of P_0 is at most κ, and the contraction inequality for total variation gives

$$\mathrm{TV}(P_0^k(x, \cdot), \mu) \leq \kappa^k. \tag{146}$$

Define its common mass and deficit by

$$a_k = \sum_y \min_x P_0^k(x, y), \qquad h_k = 1 - a_k.$$

We first prove the dimension-free recurrence

$$h_{k+1} \leq \kappa h_k + \kappa^k, \qquad h_k \leq k\kappa^{k-1} \quad (k \geq 1). \tag{147}$$

For $\mu_y > 0$, put $f_y(z) = P_0^k(z, y)/\mu_y$ and $u_y = \min_z f_y(z)$. Stationarity gives $\mu f_y = 1$, so $u_y \leq 1$. For any source row $p = P_0(x, \cdot)$, let $v = (\mu - p)_+$. Then $v \leq \mu$ coordinatewise, $v\mathbf{1} \leq \kappa$, and

$$\begin{aligned} p f_y &\geq 1 - v(f_y - u_y \mathbf{1}) \\ &\geq 1 - \mu(f_y - \mathbf{1})_+ - \kappa(1 - u_y). \end{aligned}$$

The first inequality uses $p - \mu = (p - \mu)_+ - v$ and the nonnegativity of $f_y - u_y\mathbf{1}$. For the second, use $f_y - u_y\mathbf{1} \leq (f_y - \mathbf{1})_+ + (1 - u_y)\mathbf{1}$. After multiplication by μ_y and summation over y,

$$\sum_{y:\mu_y > 0} \mu_y \mu (f_y - \mathbf{1})_+ = \sum_z \mu_z \, \mathrm{TV}(P_0^k(z, \cdot), \mu) \leq \kappa^k.$$

Zero stationary coordinates cause no exception: if $\mu_y = 0$, then $P_0^k(z, y) = 0$ for every z in the support of μ, and hence $\min_x P_0^k(x, y) = 0$. Taking the minimum over x in the preceding inequality therefore gives the recurrence for all states, including transient ones. Its displayed consequence follows by induction from $h_1 \leq 1$.

Let $K = K_\kappa$, $\beta = \beta_\kappa$, and $\eta = \eta_\kappa$ be as in (132). Then $a_K \geq \beta$. Let $w_y = \min_x P_0^K(x, y)$ and set $\nu = w/a_K$. There are stochastic rows R_x with

$$P_0^K(x, \cdot) = \beta\nu + (1 - \beta)R_x. \tag{148}$$

Indeed $P_0^K(x, \cdot) \geq a_K\nu \geq \beta\nu$. Stationarity also gives $\mu \geq a_K\nu$, so $\mathrm{TV}(\mu, \nu) \leq 1 - a_K \leq (1 - \kappa)/4$.

For each goal g, choose a stationary minimizing policy with kernel P_g and an initial state s_g attaining m_g. Every row of P_g^K is a convex combination of legal one-step rows, as is $\mu = \mu P_0$. The global diameter assumption extends to this convex hull. Consequently, for every target state y,

$$\mathrm{TV}(\nu, P_g^K(y, \cdot)) \leq \frac{1 - \kappa}{4} + \kappa = 1 - \eta. \tag{149}$$

No common policy or stationary law is required for the targets.

Start the source X at the state s whose centered row is being bounded, and target Y^g at s_g. Divide time into blocks of K transitions. At block k, draw a fresh common coin $B_k \sim \mathrm{Bernoulli}(\beta)$, independent of the joint past. Conditional on $B_k = 1$, draw the source endpoint from ν; conditional on $B_k = 0$, draw it from R_x, where x is the current source state. On success, maximally couple this endpoint with each as-yet unmatched target endpoint of law $P_g^K(y_g, \cdot)$. All pair couplings can share one source draw: sample each target from its conditional law given that draw. On failure, draw target endpoints independently with the same required marginal. Targets that have already met may continue independently under their own policies. Given the endpoints, fill in each component's block by its ordinary K-step Markov bridge.

The source's complete block path, after averaging the fresh coin, has exactly the P_0 path law conditional on the preceding joint boundary history. Every target's complete block path has exactly its P_g path law, even conditional on the fresh coin. Only bridges with positive endpoint probability are needed. Let T_g be the first successful block endpoint at which X and Y^g agree. Follow the source until T_g and target g thereafter to define a spliced process Z^g.

This splice is a legal randomized policy. At a block boundary its mode is determined by the past. In source mode the next entire block has the ordinary P_0 path law; in target mode it has the ordinary P_g path law. Hence conditioning further on a prefix of the *selected* block gives the correct physical transition row. Averaging over boundary histories compatible with the observed states and actions preserves that row. Auxiliary endpoints and bridges are not disclosed before the end of their block. Equivalently, first execute the selected component's ordinary block, then sample the auxiliary variables from their conditional joint distribution given that completed block. Thus no action uses a future meeting test.

The target has expected goal cost m_g, and optimality bounds the cost of Z^g below by M_{sg}. Their costs coincide from T_g onward. Dropping the nonpositive target cost and then the discount gives

$$D_{sg} \le \mathbb{E} \sum_{t<T_g} \gamma^t \big(\mathbf{1}\{X_t = g\} - \mathbf{1}\{Y_t^g = g\} \big) \le \mathbb{E} \sum_{t<T_g} \mathbf{1}\{X_t = g\}. \tag{150}$$

For the joint bound, condition on the entire coin sequence B. Let J_r be the index of its r-th success and $t_r = KJ_r$. Let A_g^r be the event that g is unmatched at the *start* of block J_r. Each preceding successful block has conditional meeting probability at least η, uniformly in its joint boundary history, by (149). Future coins are independent of all preceding endpoint and bridge draws. Induction therefore gives

$$\Pr(A_g^r \mid B) \le (1-\eta)^{r-1}. \tag{151}$$

At the endpoint of block J_r, the source has law ν independent of its preceding joint history, including A_g^r, conditional on B. Its future endpoint and bridge laws depend only on its own current state and the prescribed coins. It follows that, for each $t_r \le t < t_{r+1}$, X_t and A_g^r are independent conditional on B. This independence concerns the old unmatched set; the set updated at t_r may depend on the new source state.

Since $t < T_g$ in this interval implies A_g^r, we obtain

$$\begin{aligned} \mathbb{E}\left[\sum_g \mathbf{1}\{X_t = g, t < T_g\} \,\middle|\, B \right] &\le \sum_g \Pr(X_t = g \mid B) \Pr(A_g^r \mid B) \\ &\le (1-\eta)^{r-1}. \end{aligned} \tag{152}$$

Before t_1 the same sum is at most one, since the source occupies one state. Almost surely there are infinitely many successes, and $\mathbb{E}t_1 = \mathbb{E}(t_{r+1} - t_r) = K/\beta$. Sum (152) over its intervals before averaging over the coins. Tonelli's theorem and (150) give

$$\sum_g D_{sg} \le \frac{K}{\beta} + \sum_{r \ge 1} \frac{K}{\beta} (1-\eta)^{r-1} = \frac{K}{\beta} \left(1 + \frac{1}{\eta} \right) = L_\kappa. \tag{153}$$

The first interval and the exponent $r-1$ account for the one-cycle delay. No independence between a cycle length and the newly updated unmatched set was used. Since s was arbitrary, this proves

$$\|D\|_\infty \le L_\kappa. \tag{154}$$

When $\kappa \le 1/2$, rerun the same construction with block length 6, $\beta = 13/16$, and $\eta = 5/16$, forming ν from the six-step common part. The recurrence gives $a_6 \ge 13/16$, while $\mathrm{TV}(\nu, P_g^6(y,\cdot)) \le 3/16 + 1/2 = 11/16$. The resulting series is $2016/65 < L_0 = 32$, proving the stated special case without changing the constants in Theorem 5. □

For the remaining argument, which proves the diameter-$1/2$ sequential theorem, set $L = L_0$. The budget also controls the total column minimum at the relevant inverse-mass scale. Indeed,

$$d^\top D = m^\top CM - (d^\top \mathbf{1})m^\top = t_f m^\top.$$

Together with $d \ge 0$, $d^\top \mathbf{1} < 1$, and $t_f \ge c_*/2$ from (145), this gives

$$t_f m^\top \mathbf{1} = d^\top D \mathbf{1} \le L, \qquad m^\top \mathbf{1} \le \frac{2L}{c_*}. \tag{155}$$

The factor c_*^{-1} will cancel when we restrict to labels whose inverse mass is small.

E.3 Overlap leaves inverse mass outside the initial batch

Lemma 3 (Inverse mass outside an initial batch)**.** *Suppose $0 < \kappa < 1$ and $\kappa(P) \le \kappa$. Let I_0 contain m labels, $1 \le m \le n-1$, and put $\rho = 1 - (m-1)\kappa > 0$. Then*

$$\sum_{g \notin I_0} c_g \ge \eta^+_{\kappa,m} S_c, \qquad \eta_{\kappa,m} := \eta^+_{\kappa,m} = \min\left\{ \frac{1}{2m+1}, \frac{\rho(1-\kappa)}{6(m+1)} \right\}. \tag{156}$$

If $n \ge 4$ and $\kappa(P) \le 1/2$, every two-label set also satisfies the sharper bound

$$\sum_{g \notin I_0} c_g \ge \frac{S_c}{16}. \tag{157}$$

Proof. Put $v = c/c_*$, $s_0 = \sum_{g \notin I_0} v_g$, $q = n - m \ge 1$, and $S_{\gamma,\kappa} = (1 - \gamma\kappa)^{-1}$. There is a scalar $\lambda > 0$ with $Dv = \lambda \mathbf{1}$. At a label with $v_g = 1$, (142) gives $\lambda \ge 1/(1+\gamma\kappa)$. Let $\varrho_g = a m_g$ and $R = \sum_g v_g \varrho_g$. The centered Bellman equations read

$$D_g(s) = \mathbf{1}\{s = g\} - \varrho_g + \gamma \min_u P_{su} D_g.$$

A sum of minima is at most the minimum of the sum. Multiplication by v_g and summation therefore gives $R + a\lambda \le v_s$ for every state. Thus $R \le \min_{s \notin I_0} v_s \le s_0/q$.

For each goal in I_0, choose a minimizing legal row p_g at a state where its centered column is zero. Their coordinatewise minimum ν_0 has mass at least ρ: relative to one of these rows, taking successive minima loses at most the sum of its TV distances to the other $m-1$ rows. Thus

$$\gamma \nu_0 \sum_{g \in I_0} v_g D_g \le \sum_{g \in I_0} v_g \varrho_g \le R \le s_0/q.$$

Since every centered entry is at most $S_{\gamma,\kappa}$, the vector inside ν_0 is nonnegative and at least $(\lambda - S_{\gamma,\kappa} s_0)\mathbf{1}$. If that scalar is negative, nonnegativity still gives the following lower bound. Hence

$$\frac{s_0}{q} \ge \gamma\rho(\lambda - S_{\gamma,\kappa} s_0), \qquad s_0 \ge \frac{\gamma\rho\lambda}{q^{-1} + \gamma\rho S_{\gamma,\kappa}}.$$

For $\gamma \ge 1/2$, we have $\gamma\lambda \ge 1/3$ and $\gamma S_{\gamma,\kappa} \le (1-\kappa)^{-1}$. Since $q^{-1} \le 1$,

$$s_0 \ge \frac{\rho(1-\kappa)}{3(1-\kappa+\rho)} \ge \frac{\rho(1-\kappa)}{6} =: x.$$

The inside normalized mass is at most m, and $0 < x \le 1$, so the outside fraction is at least $x/(m+x) \ge x/(m+1)$. For $\gamma < 1/2$, every $c_g \ge 1 - \gamma > 1/2$; the outside fraction exceeds $q/(2m+q) \ge 1/(2m+1)$. Taking the smaller of these two guarantees proves (156) and covers $\gamma = 0$.

For a pair under $\kappa \le 1/2$, retain $q \ge 2$ and $\rho \ge 1/2$ in the displayed pre-specialization bound. At $\gamma \ge 1/2$, use $\lambda \ge (1+\gamma/2)^{-1}$ and $S_{\gamma,\kappa} \le (1-\gamma/2)^{-1}$ to get $s_0 \ge 2/15$, and hence the fraction $(2/15)/(2+2/15) = 1/16$. At $\gamma < 1/2$ the outside fraction exceeds $1/3$. □

E.4 Interpolation identities and score perturbation

Fix a nonempty proper acquired set I, with complement U. Every principal block of C is invertible with nonnegative inverse: write $C_{UU} = I - Q$, where $Q \geq 0$ and $Q\mathbf{1} < \mathbf{1}$, and apply the Neumann series. Schur complementation gives

$$B := M_{II}^{-1} = C_{II} - C_{IU}C_{UU}^{-1}C_{UI}. \tag{158}$$

The subtracted matrix is nonnegative. Thus B has nonpositive off-diagonal entries and diagonal at most one. Its row sums are $c_I - C_{IU}C_{UU}^{-1}c_U > 0$, so

$$\|B\|_\infty \leq 2. \tag{159}$$

Define

$$x = B\mathbf{1}_I, \qquad X = M_{UI}B, \qquad b = \mathbf{1} - M_{:I}x. \tag{160}$$

Block inversion gives

$$\begin{aligned} X = -C_{UU}^{-1}C_{UI} \geq 0, \quad b_I = 0, \qquad\qquad b_U = C_{UU}^{-1}c_U > 0, \\ X\mathbf{1}_I = \mathbf{1}_U - b_U \leq \mathbf{1}_U, \quad x = c_I - C_{IU}b_U > 0. \end{aligned} \tag{161}$$

In particular, $b_U \leq \mathbf{1}_U$. The off-diagonal absolute row sums of C are at most one, so

$$\|x\|_1 \leq S_c + |I| \max_{s \in U} b_s. \tag{162}$$

For $g \in U$ and nonempty $V = U \setminus \{g\}$, the V rows of $C_{UU}b_U = c_U$ give

$$b_V = C_{VV}^{-1}c_V + qb_g, \qquad q = -C_{VV}^{-1}C_{Vg}, \qquad 0 \leq q \leq \mathbf{1}_V. \tag{163}$$

The last inequality follows by multiplying $C_{VV}\mathbf{1}_V + C_{Vg} = c_V - C_{VI}\mathbf{1}_I \geq 0$ by C_{VV}^{-1}. The g row also gives

$$b_g \geq c_g / C_{gg} \geq c_g. \tag{164}$$

Now let $\widetilde{M} = M + E$ with $\|E\|_{\max} \leq \varepsilon$, and write $k = |I|$. We use matrix one-norms (maximum absolute column sums) to retain $\|x\|_1$ rather than $k\|x\|_\infty$. Every column of BE_{II} has absolute sum at most

$$\varepsilon \sum_{i,h \in I} |B_{ih}| \leq 2k\varepsilon.$$

If $2k\varepsilon \leq 1/2$, then $I + BE_{II}$ and $\widetilde{M}_{II}$ are invertible, and

$$\widehat{x} = \widetilde{M}_{II}^{-1}\mathbf{1}_I = (I + BE_{II})^{-1}x, \qquad \|\widehat{x}\|_1 \leq 2\|x\|_1. \tag{165}$$

Since $\widehat{x} - x = -BE_{II}\widehat{x}$, the uncentered entries cancel from the score error:

$$\begin{aligned} \widetilde{b}_U - b_U = (XE_{II} - E_{UI})\widehat{x}, \\ \|\widetilde{b}_U - b_U\|_\infty \leq 2\varepsilon\|\widehat{x}\|_1 \leq 4\varepsilon\|x\|_1. \end{aligned} \tag{166}$$

Here $X \geq 0$ and $X\mathbf{1} \leq \mathbf{1}$ imply that every entry of $XE_{II} - E_{UI}$ has absolute value at most 2ε. The estimate holds for every acquired set simultaneously. In particular, it applies to the sets selected using this same perturbed matrix, without conditioning on an exact-input query path.

E.5 Preserving average inverse mass across all scales

We give the proof for any initial set I_0 whose outside mass is at least ηS_c, where $0 < \eta \leq 1$. Set

$$L = L_0 = 32, \qquad \alpha = 2^{-8L}\min\{\eta, (32L)^{-1}\}, \qquad \varepsilon_0 = \frac{\alpha}{64n}. \tag{167}$$

The pair bound uses $\eta = 1/16$, recovering $\alpha = \alpha_0$.

Proof of Lemma 2. Assume $\|\widetilde{M} - M\|_{\max} \leq \varepsilon_0$. Every noisy principal block of size at most $n - 1$ is invertible, since $2n\varepsilon_0 = \alpha/32 < 1/2$. Thus no singular-block fallback occurs.

Fix the realized query path. Some label outside I_0 has mass at least $\eta S_c/n$. If $|I_0| = n-1$, the sole omitted label already has mass at least ηS_c, and the claim follows. Otherwise define the absolute thresholds

$$\theta_0 = \min\left\{\frac{\eta S_c}{n}, \frac{c_*}{32L}\right\}, \qquad z_0 = \theta_0 2^{-8L}. \tag{168}$$

Since $S_c \le n c_*$ and $L \ge 1$,

$$\theta_0 \ge \min\{\eta, (32L)^{-1}\}\frac{S_c}{n}, \qquad z_0 \ge \alpha\frac{S_c}{n}. \tag{169}$$

Only the analysis uses these quantities.

Let j be the final omitted label, and suppose $c_j < z_0$. For each $\theta \in [z_0, \theta_0]$, there is a label outside I_0 with mass at least θ. Because j has smaller mass, there is a last such label queried in the suffix. Call it g; let I be the acquired set just before that query, $U = I^c$, and $V = U \setminus \{g\}$ the set left afterward. Then

$$j \in V, \qquad c_l < \theta \ (l \in V), \qquad b_g \ge c_g \ge \theta. \tag{170}$$

Put $A = C_{VV}^{-1} \ge 0$ and $u = Ac_V$. Restrict (143) to $V \times V$ and multiply by A to get

$$A = D_{VV} + AC_{V,V^c}D_{V^c,V} + um_V^\top \le D_{VV} + um_V^\top.$$

The cross-block term is nonpositive. By (155) and the upper limit in (168),

$$\omega := m_V^\top c_V \le \frac{2L\theta}{c_*} \le \frac{1}{16}.$$

Multiplication by c_V gives

$$(1-\omega)u \le D_{VV}c_V, \qquad \|u\|_\infty \le \frac{L\theta}{1-\omega} \le 2L\theta. \tag{171}$$

The residual identity (163) says $b_V = u + qb_g$ with $q \le \mathbf{1}$. Together with (162), this yields

$$\max_U b \le b_g + 2L\theta, \qquad \|x\|_1 \le S_c + n(b_g + 2L\theta). \tag{172}$$

Since the rule selected g, its noisy score satisfies $\widetilde{b}_g \le \widetilde{b}_j$. Apply (166) to both coordinates and use $b_g \ge \theta \ge z_0 \ge \alpha S_c/n$:

$$\begin{aligned}\frac{b_g - b_j}{b_g} &\le 8\varepsilon_0\left[\frac{S_c}{\theta} + n(1+2L)\right] \\ &\le 8n\varepsilon_0[\alpha^{-1} + 1 + 2L] = \frac{1}{8}[1 + \alpha(1+2L)] \le \frac{35}{256} < \frac{1}{4}.\end{aligned} \tag{173}$$

For the last inequality, use $\alpha \le 1/(32L)$ and $1 + 2L \le 3L$. Thus $b_j \ge (3/4)b_g$ on the actual noisy path.

We next use column shape to sharpen the coarse bound $q \le \mathbf{1}$. Column g of (143) gives

$$q\Delta_g = D_{Vg} + AC_{VI}D_{Ig} + um_g \le D_{Vg} + um_g.$$

Hence $q_j \le 1/2 + u_j m_g/\Delta_g$. The bounds $m_g/\Delta_g \le 4L/c_*$ and $b_j = u_j + q_j b_g \ge (3/4)b_g$ imply

$$\frac{1}{4}b_g \le u_j(1 + 4Lb_g/c_*).$$

The function $y/(1 + 4Ly/c_*)$ is increasing. Since $b_g \ge \theta$ and $\theta \le c_*/(32L)$,

$$u_j \ge \frac{\theta}{4(1 + 4L\theta/c_*)} \ge \frac{2}{9}\theta.$$

Applying (171) at j now gives

$$\sum_{l \in V} D_{jl}c_l \ge (1-\omega)u_j \ge \frac{5}{24}\theta \ge \frac{\theta}{5}. \tag{174}$$

For this fixed final row j, define

$$F_j(\theta) = \sum_{l:c_l<\theta} D_{jl}c_l.$$

Each set V in (170) is contained in the corresponding threshold set. Therefore (174) implies $F_j(\theta) \geq \theta/5$ throughout $[z_0, \theta_0]$. Although different thresholds may refer to different query stages, the lower bounds concern a single function. Integrating once gives

$$\begin{aligned}\frac{1}{5}\log\frac{\theta_0}{z_0} &\leq \int_{z_0}^{\theta_0} \frac{F_j(\theta)}{\theta^2}\,d\theta \\ &= \sum_{l:c_l<\theta_0} D_{jl}c_l\left(\frac{1}{\max\{z_0, c_l\}} - \frac{1}{\theta_0}\right) \leq \sum_l D_{jl} \leq L.\end{aligned} \tag{175}$$

Each centered occupation term is charged at most once. But the left side equals $(8\log 2/5)L > L$, a contradiction. Ties among inverse masses change only finitely many integration points. Thus $c_j \geq z_0 \geq \alpha S_c/n$. Taking I_0 to be the prescribed or actual first pair and $\eta = 1/16$ proves the stated lemma. The argument also proves the same conclusion for any initial set satisfying the outside-mass bound, with (167). □

E.6 Decoding the retained table

Proof of Theorem 5. For this theorem use $\eta = 1/16$ and $\alpha = \alpha_0$ in (167). Let the complete normalized input satisfy $\|Z - \bar{q}_P\|_{\max} \leq \delta$, and put $\varepsilon_q = H\delta$. For each acquired goal form

$$\widetilde{q}_g(s,u) = HZ_g(s,u), \qquad \widetilde{M}_{sg} = \mathbf{1}\{s=g\} + \gamma \min_u \widetilde{q}_g(s,u).$$

The minimum is 1-Lipschitz in the maximum norm, so

$$\|\widetilde{M} - M\|_{\max} \leq \gamma\varepsilon_q. \tag{176}$$

As before, the complete perturbed matrix is a proof device; only received columns are formed by the algorithm.

First suppose $\varepsilon_q \leq \varepsilon_0$. Lemma 2 applies, because the feature error in (176) is at most ε_0. Write $J = S \setminus \{j\}$ for the retained set, so $c_j \geq \alpha S_c/n$. For any zero-sum row ξ, put $y = \xi M$. Since $yc = 0$,

$$\begin{aligned}\xi &= \sum_{g\in J} y_g\left(C_g - \frac{c_g}{c_j}C_j\right), \\ \|\xi\|_1 &\leq K_j\|(\xi M)_J\|_\infty, \qquad K_j = 2\left[(n-2) + \frac{S_c}{c_j}\right] \leq \frac{4n}{\alpha}.\end{aligned} \tag{177}$$

The bound follows by summing $\|C_g\|_1 + (c_g/c_j)\|C_j\|_1$.

For each state–action pair choose the simplex residual fit, with fixed deterministic ties:

$$\widehat{p}_{su} \in \arg\min_{z\geq 0,\ z\mathbf{1}=1} \|z\widetilde{M}_J - \widetilde{q}_J(s,u,\cdot)\|_\infty. \tag{178}$$

Write $p_{su} = P_{su}$ for the true transition row. It is feasible and has residual at most $(1+\gamma)\varepsilon_q$, since $p_{su}M_J = q_J(s,u,\cdot)$. The fitted residual is no larger. As $\widehat{p}_{su}$ is also a probability row,

$$\begin{aligned}\|(\widehat{p}_{su} - p_{su})M_J\|_\infty &\leq \|\widehat{p}_{su}(M_J - \widetilde{M}_J)\|_\infty \\ &\quad + \|\widehat{p}_{su}\widetilde{M}_J - \widetilde{q}_J(s,u,\cdot)\|_\infty \\ &\quad + \|\widetilde{q}_J(s,u,\cdot) - q_J(s,u,\cdot)\|_\infty \\ &\leq 2(1+\gamma)\varepsilon_q \leq 4\varepsilon_q.\end{aligned} \tag{179}$$

The row difference sums to zero, so (177) bounds its norm by $4K_j\varepsilon_q \leq 16n\varepsilon_q/\alpha$.

For $\varepsilon_q > \varepsilon_0$, the simplex fit still exists: its objective is continuous on a compact simplex, even if the observed matrix is singular. Its output is stochastic, so the recovery error is at most two. Moreover

$$2 < \frac{2}{\varepsilon_0}\varepsilon_q = \frac{128n}{\alpha}\varepsilon_q.$$

Combining the two regimes proves

$$\max_{s,u} \|\widehat{p}_{su} - p_{su}\|_1 \le \min\left\{2, \frac{128n}{\alpha} H\delta\right\}. \tag{180}$$

This reasoning includes $\gamma = 0$, when the constructed feature columns are exact coordinate vectors. Neither acquisition nor decoding uses δ or the proof thresholds.

Substituting $L = 32$ and $\eta = 1/16$ in (167) gives

$$\alpha = 2^{-266}, \qquad \frac{128}{\alpha} = 2^{273} = C_{\mathrm{rec}}. \tag{181}$$

This is an absolute coefficient; no numerical sharpness is asserted. For a general initial set in Lemma 3, the same decoding argument gives (180) with $\alpha = 2^{-256} \min\{\eta^+_{\kappa,m}, 2^{-10}\}$. □

E.7 A MATCHING RECOVERY LOWER BOUND ON THE SAME CLASS

We prove the comparison in (187) for $n \ge 4$ and at least four action labels per state. Let $N = 2\lfloor n/2 \rfloor$. Choose a three-state core and let w be uniform on it. At every state, make available the three actions uniform on the three different pairs of core states. Their pairwise total-variation distances equal $1/2$. For each singleton goal, one of these actions avoids it, and it can be used at every state thereafter. Thus the minimum continuation cost is zero and $M = I$. Any additional actions preserve this identity because all costs are nonnegative.

Choose N states and give them balanced signs $\sigma_i \in \{-1, 1\}$, so $\sum_i \sigma_i = 0$. For $t \in [0, 1]$, define probability rows $v_\pm$, zero outside these states, and their mixtures with w by

$$v_\pm(i) = \frac{1 \pm t\sigma_i}{N}, \qquad p_\pm = \frac{3}{4} w + \frac{1}{4} v_\pm.$$

At one distinguished source state o, use p_+ or p_- as a fourth action in two environments P_+ and P_-, respectively. At all other states the fourth action duplicates a core action; any further labels do so as well. Each probe row has mass at least $1/4$ at each core state. Its overlap with every pair-of-core row is therefore at least $1/2$. Hence the global diameter in each environment is at most $1/2$.

The two environments have $M = I$ and differ only in the probe row. Their entire raw arrival-critic tables equal their transition tables: $q_\pm = P_\pm$. In particular,

$$\|p_+ - p_-\|_1 = \frac{t}{2}, \qquad \|\bar{q}_+ - \bar{q}_-\|_{\max} = \frac{t}{2NH}. \tag{182}$$

Take $t = \min\{1, 4NH\delta\}$. The single complete table $Z = (\bar{q}_+ + \bar{q}_-)/2$ lies within entrywise distance δ of both true normalized tables. It is fixed before queries or private randomness are generated.

A full-input randomized decoder has the same output distribution in both environments when given Z. Couple its private seed. For each output, the sum of its two probe-row errors is at least $t/2$, by the triangle inequality. Taking expectations and then the worse of the two fixed environments proves

$$\mathcal{R}_{\mathrm{full}}(\delta) \ge \frac{t}{4} = \min\left\{\frac{1}{4}, NH\delta\right\}. \tag{183}$$

This also bounds expected maximum-row error. The environment is not selected after observing the decoder's seed.

A full-input procedure can simulate any permitted sequential procedure by revealing only its requested columns, including any prescribed first pair. The common complete table makes every such adaptive transcript identical under the two environments for a coupled seed. Consequently

$$\min\left\{\frac{1}{4}, 2\lfloor n/2 \rfloor H\delta\right\} \le \mathcal{R}_{\mathrm{full}}(\delta) \le \mathcal{R}_{\mathrm{seq}}(\delta) \le \min\{2, C_{\mathrm{rec}} nH\delta\}. \tag{184}$$

Here both risks use worst-case expected maximum transition-row ℓ_1 error. Since $2\lfloor n/2 \rfloor \ge 3n/4$ for $n \ge 4$, the lower bound is at least $\frac{1}{4}\min\{1, nH\delta\}$. The two recovery risks therefore have order $\Theta(\min\{1, nH\delta\})$, with absolute constants. This comparison concerns recovery from critic queries. The control lower bounds with subsequent transition observations are proved separately in Appendix G.

F Optimal recovery and the timing of feedback

The relevant measure of overlap is the global transition-row diameter

$$\kappa(P) = \max_{s,u,t,v} \mathrm{TV}(P_{su}, P_{tv}) \le \tfrac{1}{2}. \tag{185}$$

It compares all state-action rows and contracts value spans (Gaubert & Qu, 2015). Our first result asks whether leaving one column unseen costs recovery precision within this class.

Let $\mathcal{R}_{\mathrm{full}}(\delta)$ and $\mathcal{R}_{\mathrm{seq}}(\delta)$ be the smallest worst-case expected errors $\mathbb{E}\max_{s,u}\|\widehat{P}_{su} - P_{su}\|_1$ for complete observation and sequential acquisition, respectively. The sequential procedure uses $n-1$ queries in total, with any prescribed initial pair. In both risks, the kernel and complete noisy table are fixed before private randomness.

Theorem 5 (Optimal recovery at the minimum query budget)**.** *Let $n \ge 4$, $H \ge 1$, and $\kappa(P) \le 1/2$, with finite nonempty action sets. The sequential rule in Section D, starting from the empty set or any prescribed pair, and the fit in* (129) *satisfy*

$$\max_{s,u} \|\widehat{P}_{su} - P_{su}\|_1 \le \min\{2, C_{\mathrm{rec}} n H \delta\} \tag{186}$$

for an absolute constant C_{rec} and every table satisfying (3)*. With a common action count $A \ge 4$, this rate is optimal even for complete observation:*

$$\mathcal{R}_{\mathrm{full}}(\delta) \asymp \mathcal{R}_{\mathrm{seq}}(\delta) \asymp \min\{1, nH\delta\}, \tag{187}$$

with absolute comparison constants, uniformly in n, H, δ.

The matching lower bound uses two admissible kernels whose minimum-occupation matrices both equal the identity. Their complete critic tables admit a common noisy input although one transition row differs by an amount of order $\min\{1, nH\delta\}$ (Appendix E.7). Thus the minimum identifying budget also attains the complete library's optimal recovery order. Acquiring all n columns and then choosing $n-1$ to retain has a linear guarantee as well (Corollary 7). The sequential rule achieves the same order without reading the column it ultimately omits.

The second result determines how many queries may precede the first response. Write $\mathcal{D}_m^{\mathrm{batch}}(\kappa)$ for the minimax execution risk at total budget $n-1$ when the first m labels must be chosen before any reply, with the kernel class restricted to $\kappa(P) \le \kappa$. Appendix G gives the formal definition.

Theorem 6 (Overlap determines the initial commitment boundary)**.** *Fix $0 < \kappa \le 1/2$ and put $b = \lceil 1/\kappa \rceil$. For a kernel with $\kappa(P) \le \kappa$, $n \ge 4$, finite nonempty action sets, and any prescribed $1 \le m \le \min\{b, n-1\}$ initial labels, the sequential rule satisfies*

$$\max_{s,u} \|\widehat{P}_{su} - P_{su}\|_1 \le \min\{2, C_{\kappa,m} n H \delta\}, \tag{188}$$

where $C_{\kappa,m}$ is independent of n, H, δ. If $n \ge b+2$, $A \ge b+3$, and $H \ge 2$, then for every $0 \le m \le n-1$ the execution risk satisfies

$$\mathcal{D}_m^{\mathrm{batch}}(\kappa) = \begin{cases} \widetilde{\Theta}_{n,A,\kappa}(\min\{\delta H, H^{-1/2}\}), & 0 \le m \le b, \\ \widetilde{\Theta}_{n,A,\kappa}(\min\{\delta H^2, H^{-1/2}\}), & b+1 \le m \le n-1. \end{cases} \tag{189}$$

The total budget is $n-1$. The initial set may be randomized, and both subsequent acquisition and execution may be arbitrarily adaptive. Each displayed minimum times a positive constant depending only on n, A, κ is a lower bound; multiplying it by $C_{n,A,\kappa}\sqrt{\log(eH)}$ gives an upper bound.

The threshold follows from overlap among the initial goals' minimizing transition rows. Any m rows share at least $1-(m-1)\kappa$ mass. When this quantity is positive, the initial queries cannot exhaust all stable omissions. A core of $b+1$ labels can exhaust them. Table 3 gives three instances of this boundary. If $b \ge n-1$, every possible initial batch is covered by the recovery upper bound; the execution separation requires the larger state and action counts in the theorem. Constants need not remain bounded as $1-(m-1)\kappa$ approaches zero.

For diameter $1/2$, any two initial labels preserve the full-library execution rate, whereas three can incur an extra horizon factor. At $\delta \asymp H^{-2}$, the two rates are H^{-1} and $H^{-1/2}$, up to logarithms.

Table 3: **Stronger overlap permits more queries before feedback.** Consequences of Theorem 6, at total budget $n-1$. The first column bounds $\kappa(P)$; the last column gives sufficient state and action counts for the matching execution separation.

Row diameter	Largest safe initial batch	First potentially costly initial batch	Required sizes
$1/2$	2	3	$n \geq 4,\ A \geq 5$
$1/3$	3	4	$n \geq 5,\ A \geq 6$
$1/4$	4	5	$n \geq 6,\ A \geq 7$

Learning during execution limits the second loss, but does not close the gap. At $\delta = 0$, every $n-1$ subset identifies the kernel and both losses are zero. On the diameter-$1/2$ class, at most $n-2$ queries can instead leave execution risk of order $\widetilde{\Theta}_{n,A}(H^{-1/2})$ even with exact critics (Appendix G).

Selection and fitting use the same noisy table and need neither the downstream reward nor δ. The execution rule uses the declared error budget to choose between fitted-model planning and an online learner (Fruit et al., 2018). Under (185), Theorem 5 gives the upper envelope

$$\min\{1, 2C_{\mathrm{rec}} nH\delta, B(H)\}, \tag{190}$$

where $B(H)$ is the online loss bound in Appendix G, (194). Recovery is optimal in all three parameters n, H, δ; the execution comparisons fix n, A, κ.

G Query budgets, initial batches, and execution risk

We prove the execution boundary in Theorem 6 and the query-count law at diameter $1/2$. Appendix E establishes linear transition recovery after a permitted initial batch. Here we transfer that guarantee to control and prove matching fixed-dimension lower bounds for arbitrary history-dependent execution.

G.1 Admissible algorithms and quantitative envelopes

Fix $n \geq 4$, a common action count $A \geq 5$, $H \geq 2$, and $\delta \geq 0$. Put $a = H^{-1} = 1 - \gamma$ and

$$\begin{aligned} \mathcal{I}_\delta = \{(P, Z, r, o) : \kappa(P) \leq 1/2, \quad \|Z - \bar{q}_{P,\mathrm{full}}\|_{\max} \leq \delta, \\ r \in [0,1]^n, \quad o \in S\}. \end{aligned} \tag{191}$$

The kernel, complete table Z, reward, and root are fixed before the learner's private randomness. A query returns an entire column Z_g, including every state–action entry. Acquisition uses no transition observations and finishes before r, o are disclosed. Execution may use the entire subsequent evaluated trajectory, with no preliminary interaction. The dimensions, labels, discount, and declared error budget are known.

Let $\mathfrak{A}_k$ be the possibly randomized acquisition and execution algorithms using at most k distinct queries. Write $\mathcal{D}_k = \mathcal{D}_k^{\mathrm{ad}}$ for their minimax risk:

$$\mathcal{D}_k(\delta; H) = \inf_{\mathsf{A} \in \mathfrak{A}_k} \sup_{(P,Z,r,o) \in \mathcal{I}_\delta} a \left\{ J_P^*(o) - \mathbb{E}_{P,Z,r,o}^{\mathsf{A}} \left[\sum_{t \geq 0} \gamma^t r(S_{t+1}) \right] \right\}. \tag{192}$$

The expectation includes the private stream and the trajectory. For $0 \leq m \leq n-1$, let $\mathfrak{B}_m \subseteq \mathfrak{A}_{n-1}$ require the first m distinct labels to be chosen before any responses. These labels may be randomized. After receiving those m columns, the suffix may use every response so far. Replacing $\mathfrak{A}_k$ by $\mathfrak{B}_m$ in (192) defines $\mathcal{D}_m^{\mathrm{batch}}$. Thus m is a required initial commitment, not an optional maximum.

Nonadaptive acquisition is the case $m = n-1$: $\mathcal{D}_{\mathrm{na}} := \mathcal{D}_{n-1}^{\mathrm{na}} = \mathcal{D}_{n-1}^{\mathrm{batch}}$. A rule using fewer predetermined queries can acquire extra columns and ignore them. A prespecified omitted label j further restricts this class; denote its risk by $\mathcal{D}_{\mathrm{del},j}$. Finally, $\mathcal{D}_n$ equals the full-table risk, since a full-table algorithm can simulate any sequence of queries.

Use the absolute recovery coefficient from Theorem 5:

$$C_{\mathrm{rec}} = 2^{273}, \qquad \max_{s,u} \|\widehat{P}_{su} - P_{su}\|_1 \le \min\{2, C_{\mathrm{rec}} nH\delta\}. \tag{193}$$

Put $B(H) = B_{n,nA}(H)$ from Appendix L. The online rule there ignores the critics and satisfies

$$\mathrm{loss} \le B(H) \le C_{n,A}\sqrt{\log(eH)}\, H^{-1/2}. \tag{194}$$

The query-count envelopes are

$$\begin{aligned} \mathcal{D}_k(\delta; H) &\ge \begin{cases} \dfrac{H^{-1/2}}{64n\binom{n}{2}}, & 0 \le k \le n-2, \\ \min\{\delta H/5, H^{-1/2}/640\}, & n-1 \le k \le n, \end{cases} \\ \mathcal{D}_k(\delta; H) &\le \begin{cases} \min\{B(H), 1\}, & 0 \le k \le n-2, \\ \min\{1, 2C_{\mathrm{rec}} nH\delta, B(H)\}, & n-1 \le k \le n. \end{cases} \end{aligned} \tag{195}$$

For the batch envelopes, put

$$p_{n,m} = \frac{\binom{m}{3}}{\binom{n}{3}(n-3)} \qquad (3 \le m \le n-1). \tag{196}$$

Then

$$\begin{aligned} \mathcal{D}^{\mathrm{batch}}_m(\delta; H) &\ge \begin{cases} \min\{\delta H/5, H^{-1/2}/640\}, & 0 \le m \le 2, \\ p_{n,m} \min\{3\delta H^2/5, H^{-1/2}/640\}, & 3 \le m \le n-1, \end{cases} \\ \mathcal{D}^{\mathrm{batch}}_m(\delta; H) &\le \begin{cases} \min\{1, 2C_{\mathrm{rec}} nH\delta, B(H)\}, & 0 \le m \le 2, \\ \min\{8(n-1)\delta H^2, B(H), 1\}, & 3 \le m \le n-1. \end{cases} \end{aligned} \tag{197}$$

For nonadaptive acquisition, $p_{n,n-1} = 1/n$. A fixed omission j has the same upper envelope and the lower without this factor.

For fixed n, A, these bounds imply

$$\begin{aligned} \mathcal{D}_k &= \widetilde{\Theta}_{n,A}(H^{-1/2}), & k \le n-2, \\ \mathcal{D}_k &= \widetilde{\Theta}_{n,A}(\min\{\delta H, H^{-1/2}\}), & n-1 \le k \le n, \\ \mathcal{D}^{\mathrm{batch}}_m &= \widetilde{\Theta}_{n,A}(\min\{\delta H, H^{-1/2}\}), & m \le 2, \\ \mathcal{D}^{\mathrm{batch}}_m &= \widetilde{\Theta}_{n,A}(\min\{\delta H^2, H^{-1/2}\}), & 3 \le m \le n-1. \end{aligned} \tag{198}$$

The upper and lower comparisons differ by at most constants depending on n, A and the factor $\sqrt{\log(eH)}$. The linear upper coefficient in (193) does not assert matching dimension dependence for these execution lower bounds. The separate recovery comparison appears in (187).

G.2 From the fitted kernel to control

Suppose acquisition returns a stochastic kernel $\widehat{P}$ with maximum row ℓ_1 error at most e. For any stationary policy π, the true current-state value satisfies

$$\mathrm{span}(W^\pi_P) \le \frac{1}{1-\gamma/2}. \tag{199}$$

Its transition kernel inherits the global diameter bound, so this follows by applying span contraction to its Bellman equation. Differences of probability rows cancel constants, giving

$$\|(P_\pi - \widehat{P}_\pi)W^\pi_P\|_\infty \le \frac{e}{2(1-\gamma/2)}.$$

The arrival-value comparison is

$$J^\pi_P - J^\pi_{\widehat{P}} = (I - \gamma\widehat{P}_\pi)^{-1}(P_\pi - \widehat{P}_\pi)W^\pi_P. \tag{200}$$

The resolvent has infinity norm H, canceled by normalization. Comparing a fitted-model optimal policy $\widehat{\pi}$ and a true optimal policy therefore gives

$$\max_o a\{J^*_P(o) - J^{\widehat{\pi}}_P(o)\} \le \frac{e}{1-\gamma/2} \le 2e. \tag{201}$$

Only the true model requires contraction. The queries and fit are reward-free, while the planned policy may depend on the revealed reward.

Theorem 5 gives the critic upper $2C_{\mathrm{rec}}nH\delta$ with $n-1$ queries after any two predetermined distinct labels. This protocol is allowed for each $m = 0, 1, 2$: when fewer early commitments are required, it may still choose its first two labels in advance. Acquisition and fitting use restrictions of the same table Z fixed before private randomness. For every larger m, a fixed set of $n-1$ retained labels is allowed: commit to m of them and acquire the rest afterward. The fixed-omission planner in Appendix L, at zero common minorization, has loss at most $8(n-1)\delta H^2$. The online rule may instead be used after any required queries. Comparing these explicit budgets using the declared δ, and using the automatic normalized-loss cap of one, proves all upper envelopes. Selection and fitting themselves do not use δ.

G.3 One noisy lower family for full and partial acquisition

Choose an ℓ-label core R, where $3 \le \ell \le n-1$ and $A \ge \ell + 2$. The choice $\ell = 3$ proves the envelopes above; larger cores will give the overlap boundary below. Put $F = S \setminus R$, and choose a distinguished $j \in F$. Let $w = \mathrm{Unif}(R)$ and

$$v_g = \mathrm{Unif}(R \setminus \{g\}) \quad (g \in R), \qquad p_\theta = \theta e_j + (1-\theta)w. \tag{202}$$

Set $\theta_0 = 1/[4(\ell-1)^2]$ and $\theta_\pm = \theta_0 \pm t$. At each core state, provide the ℓ exclusion rows v_g and two probe rows. Model zero assigns probe labels to $p_{\theta_+}, p_{\theta_-}$; model one exchanges them. Write $\sigma \in \{0,1\}$ for this model bit. At every state in F, all action labels have row w. Extra actions duplicate an exclusion row on the core. Action labels are fixed ordinals; their unknown transition rows, including the ordering of the exclusions, are defined by R.

Take $0 < t \le \theta_0/(8\sqrt{H})$. Then $\theta_\pm \in [7\theta_0/8, 9\theta_0/8]$, and every row is stochastic. Distinct exclusions are distance $1/(\ell-1)$ apart, while

$$\mathrm{TV}(v_g, p_\theta) = \frac{1+(\ell-1)\theta}{\ell} \le \frac{1}{\ell-1}, \qquad \mathrm{TV}(v_g, w) = \frac{1}{\ell},$$
$$\mathrm{TV}(p_{\theta_+}, p_{\theta_-}) = 2t, \qquad \mathrm{TV}(w, p_\theta) = \theta.$$

The probe inequality follows from $\theta \le 1/(\ell-1)^2$. Thus both kernels have diameter exactly $1/(\ell-1)$ and use $\ell+2$ actions, including at the boundary of that diameter class.

The occupation matrix, common to the two bits and to all $j \in F$, is

$$M = I + \gamma \sum_{f \in F} e_f w. \tag{203}$$

For a core goal g, the exclusion v_g avoids it forever from the core; each forced state contributes γ/ℓ before returning. For a forced goal, future minimum occupation is zero. These facts verify every Bellman equation. On core probe rows the arrival critics are

$$q_j(p_\theta) = \theta, \qquad q_g(p_\theta) = \frac{1-a\theta}{\ell} \ (g \in R), \qquad q_f(p_\theta) = 0 \ (f \in F \setminus \{j\}). \tag{204}$$

For a core goal, forced source rows have critic $1/\ell$ and exclusion source rows depend only on R and the excluded label. All nonspecial forced-goal columns vanish. In particular, every core-column midpoint is independent of both the bit and the identity of j. The complete normalized tables satisfy

$$\|\bar{q}_0 - \bar{q}_1\|_{\max} = 2at, \qquad \|\bar{q}_{0,-j} - \bar{q}_{1,-j}\|_{\max} = \frac{2a^2t}{\ell}. \tag{205}$$

Set $r = e_j$ and choose root $o = \min R$ under a fixed ordering. Only $R \cup \{j\}$ is reachable. On this subsystem,

$$W^*_\sigma(j) - W^*_\sigma(s) = \frac{1}{1+\gamma\theta_+} \qquad (s \in R).$$

The larger probe is optimal at each core state. Any other action has arrival disadvantage at least $2t/(1+\gamma\theta_+) \ge 8t/5$. A probe transition is Bernoulli between j and the common law w. In either direction,

$$\mathrm{KL}(p_{\theta_+} \| p_{\theta_-}) \le \frac{4t^2}{\theta_-(1-\theta_-)} \le \frac{8t^2}{\theta_0}, \tag{206}$$

with the same bound for the reverse direction. All other rows have zero divergence between bits.

Suppose the joint law of the acquisition transcript and the entire private random stream agrees under the two bits. The adaptive chain rule through $T = \lceil H \rceil \le 2H$ decisions then bounds history KL by $16Ht^2/\theta_0 \le \theta_0/4 \le 1/64$, and total variation by $1/8$. Each decision is in the core with probability at least $3/4$. The two models have different optimal probe labels, so the testing inequality gives bit-averaged probability of a suboptimal core action at least $(3/4 - 1/8)/2 \ge 1/4$. The arrival performance identity from Appendix L applies to any history-dependent rule. Since $a \sum_{k<T} \gamma^k \ge 1/2$, its bit-averaged loss is at least

$$\frac{8t}{5} \cdot \frac{1}{4} \cdot \frac{1}{2} = \frac{t}{5}. \tag{207}$$

This calculation holds for any common private-stream distribution, including one obtained by conditioning on acquisition.

For the full-input lower, use

$$t = \min\{\delta H, \theta_0/(8\sqrt{H})\}, \qquad Z = \tfrac{1}{2}(\bar{q}_0 + \bar{q}_1).$$

Equation (205) makes this complete common table legal under either bit. Applying (207) gives, even to a rule that reads all columns,

$$\text{full-input risk on this family } \ge \min\{\delta H/5, \theta_0/(40\sqrt{H})\}. \tag{208}$$

For partial acquisition we use the larger calibration

$$t = \min\{\ell\delta H^2, \theta_0/(8\sqrt{H})\}, \qquad (Z_\sigma)_{-j} = \tfrac{1}{2}(\bar{q}_{0,-j} + \bar{q}_{1,-j}), \quad (Z_\sigma)_j = \bar{q}_{\sigma,j}. \tag{209}$$

Every complete table is legal: its retained error is $a^2 t/\ell \le \delta$ and the special column is exact. For a fixed omitted label j, the observed transcripts agree and (207) yields

$$\text{fixed-omission risk on this family } \ge \min\{\ell\delta H^2/5, \theta_0/(40\sqrt{H})\}. \tag{210}$$

The exact completion of the special column matters: a full midpoint can violate the error budget at this calibration. For $\delta = 0$ these lower bounds are zero and require no distinct bits. Taking $\ell = 3$ gives the constants in (195) and (197). The general bounds hold in any diameter class containing these same two kernels.

G.4 Why an arbitrary adaptive suffix may be too late

Fix an algorithm in $\mathfrak{B}_m$ with $m \ge \ell$. Choose R uniformly among the ℓ-label subsets, then j uniformly from $F = S \setminus R$, and an independent fair bit. This finite prior fixes the complete kernel and table (209), reward e_j, and root $\min R$ before private randomness.

Let U_{priv} include the entire private stream, and let $I_0(U_{\mathrm{priv}})$ be the required initial set of size m. Fix a stream and a core with $R \subseteq I_0(U_{\mathrm{priv}})$. Run acquisition on a response path that returns the common midpoint for core columns and zero for every forced column. Let $J_R(U_{\mathrm{priv}})$ be its queried set, of size at most $n-1$. This path is defined without j or the bit. It contains R and therefore leaves at least one forced label unqueried.

Define

$$E = \{R \subseteq I_0(U_{\mathrm{priv}}),\ j \notin J_R(U_{\mathrm{priv}})\}. \tag{211}$$

When the first condition holds, the second is exactly the event that actual acquisition never queries j. The actual and constructed paths have identical answers until the first such query, so either both query it or both avoid it. This argument permits any response-dependent suffix, including randomized stopping. The constructed full response table need not be a legal instance: on a miss event, the observed path is realized by the legal table with its unqueried special label j.

Consequently,

$$\begin{aligned} \Pr(E \mid U_{\mathrm{priv}}, R) &= \mathbf{1}\{R \subseteq I_0(U_{\mathrm{priv}})\} \frac{|F \setminus J_R(U_{\mathrm{priv}})|}{n - \ell}, \\ \Pr(E \mid U_{\mathrm{priv}}) &\ge \frac{\binom{m}{\ell}}{\binom{n}{\ell}(n-\ell)} =: p_{n,m,\ell}. \end{aligned} \tag{212}$$

Now fix R, j for which the conditioning event has positive probability. It depends on the stream and these labels, but not on the model bit. Conditional on R, j, E, the bit remains fair, the private-stream law is the same under both models, and every acquired column agrees. The revealed reward and root also agree. No environment transition was observed during acquisition, so this conditioning does not change the subsequent transition laws.

The testing argument (206)–(207) therefore gives conditional bit-averaged loss at least $t/5$. Conditional expected losses outside E are nonnegative. Averaging the finite prior yields

$$\text{prior mean loss} \geq p_{n,m,\ell}\frac{t}{5} = p_{n,m,\ell}\min\left\{\frac{\ell\delta H^2}{5}, \frac{\theta_0}{40\sqrt{H}}\right\}. \tag{213}$$

At least one fixed instance attains this loss over the learner's randomness and trajectory. Thus the instance is not selected after observing either. At $\ell = 3$ this proves the lower batch envelope, including randomized nonadaptive acquisition at $m = n - 1$.

For a rule that adapts only its last query, $m = n - 2$ and $p_{n,n-2} = (n-4)/[n(n-1)]$ when $n \geq 5$. At $n = 4$ its two-label initial batch instead has the full-precision upper bound. This initial-batch result is distinct from the total-round lower bound in Theorem 4.

G.5 Stronger overlap delays the first necessary feedback

We prove Theorem 6. Its recovery coefficient can be taken as

$$C_{\kappa,m} = \frac{128}{2^{-256}\min\{\eta_{\kappa,m}, 2^{-10}\}}, \tag{214}$$

where $\eta_{\kappa,m}$ is the fraction in Lemma 3.

Proof. The condition $m \leq b$ is equivalent to $\mu = 1 - (m-1)\kappa > 0$. Lemma 3 gives the outside fraction $\eta_{\kappa,m}$. The parameterized proof in (167)–(180) gives (188), including $m = n - 1$ when no suffix is needed. Empty starts use Theorem 5. The control transfer and online bound apply throughout the smaller diameter class. Thus the favorable upper envelope is $\min\{1, 2C_{\kappa,m}nH\delta, B(H)\}$ for $m \geq 1$. The fixed-omission and online upper bounds apply for every initial batch size, giving the other upper branch.

For both lower branches use the family in Section G.3 with $\ell = b + 1$. It has diameter $1/b \leq \kappa$ and fits the stated state and action counts. Its complete-table midpoint gives (208) inside this same smaller class, which bounds every acquisition protocol. For $m \geq b + 1$, the fixed-prior argument in Section G.4 gives (213). The event probability and θ_0 are positive constants at fixed n, A, κ. These two calibrations establish (189). □

The upper recovery bound needs no minimum action count. The larger state and action requirements permit the lower-bound core and its probes. Constants need not stay bounded as $1 - (m-1)\kappa$ approaches zero.

G.6 A direct lower bound with at most n minus two queries

The following exact-table construction needs only $n \geq 2$, $A \geq 2$, and $H \geq 2$. Thus it applies to the main class. Choose an unordered pair $\{i, j\}$ uniformly among all $\binom{n}{2}$ pairs, orient it by $i < j$, and independently choose a fair model bit σ. Define

$$u = \mathrm{Unif}(S), \qquad t = \frac{1}{16n\sqrt{H}}, \qquad p_+ = u + t(e_i - e_j), \quad p_- = u - t(e_i - e_j). \tag{215}$$

At every state, model zero assigns p_+ to action u_{A} and p_- to action u_{B}; model one swaps them. Additional actions duplicate u_{A} in both models. The root is state 1, and the reward is $r = e_i$, disclosed after acquisition. All these objects, including the true complete critic table, are fixed by the pair and bit before private randomness. Since $t \leq 1/(2n)$, the rows are stochastic, and

$$\kappa(P_\sigma) = \mathrm{TV}(p_+, p_-) = 2t \leq 1/2.$$

State independence makes every critic column explicit. For a goal g, put $\mu_g = \min_v p_v(g)$, where v indexes actions. Discounted Bellman uniqueness gives

$$M_g = e_g + \gamma H \mu_g \mathbf{1}, \qquad q_g(s,v) = p_v(g) + \gamma H \mu_g, \qquad \bar{q}_g(s,v) = a p_v(g) + \gamma \mu_g. \tag{216}$$

Indeed, the minimum continuation of the first expression is $\mu_g + \gamma H \mu_g = H \mu_g$. For $g \notin \{i,j\}$, $p_v(g) = \mu_g = 1/n$, so every entry of the normalized column is exactly $1/n$, independently of the pair and bit. The complete critic table is exact and therefore admissible for every declared $\delta \geq 0$.

Fix an arbitrary randomized algorithm with at most $n-2$ queries, and let U_{priv} include its entire private random stream. For each realization of this stream, run its acquisition decisions on the response path that returns the constant column $1/n$ to every query. Let $J(U_{\mathrm{priv}})$ be the distinct labels on that path. Define the event

$$E = \{\{i,j\} \cap J(U_{\mathrm{priv}}) = \varnothing\}. \tag{217}$$

For a fixed stream, this is exactly the event that the actual run does not query either hidden label. Before the first such query, the actual and constant-response paths coincide. Consequently

$$\Pr(E \mid U_{\mathrm{priv}}) = \frac{\binom{n-|J(U_{\mathrm{priv}})|}{2}}{\binom{n}{2}} \geq \frac{1}{\binom{n}{2}}. \tag{218}$$

The event depends on the pair and private stream, but not on the bit. Conditional on a fixed pair and E of positive probability, the bit remains fair and the conditional private-stream law is the same under both bits. The acquisition transcript also has the same law. We may reveal the full pair after acquisition, which only helps the learner; the already fixed reward and root agree between the two conditional experiments.

For either direction of comparison, the one-step transition KL is

$$\mathrm{KL}(p_+ \| p_-) = 2t \log \frac{1/n + t}{1/n - t} \leq 8nt^2. \tag{219}$$

The reverse KL is identical. The inequality follows from $\log((1+x)/(1-x)) \leq 4x$ for $0 \leq x \leq 1/2$. Every action, including the duplicates, obeys the same bound. Condition on the pair and E, and include the private stream, acquisition transcript, revealed pair, and entire execution history in the compared laws. Their initial distributions agree. The adaptive KL chain rule before any decision $k < T = \lceil H \rceil$ therefore gives

$$\mathrm{KL}(\mathbb{P}_0^{(k)} \| \mathbb{P}_1^{(k)}) \leq 16nHt^2 = \frac{1}{16n} \leq \frac{1}{32}, \qquad \mathrm{TV}(\mathbb{P}_0^{(k)}, \mathbb{P}_1^{(k)}) \leq \frac{1}{8}. \tag{220}$$

Conditioning on E has changed the private-stream distribution in the same way under both bits, so it creates no initial divergence.

Under model zero the optimal actions are u_{A} and its duplicates; under model one the unique optimal action is u_{B}. These disjoint sets exhaust the labels. If $B_\sigma(k)$ is the event of a suboptimal action under bit σ, then the testing inequality from (220) gives

$$\frac{\Pr_0(B_0(k)) + \Pr_1(B_1(k))}{2} \geq \frac{1 - \mathrm{TV}(\mathbb{P}_0^{(k)}, \mathbb{P}_1^{(k)})}{2} \geq \frac{1}{4}. \tag{221}$$

This holds for policies using every previous state, action, reward, and private random choice.

Here a suboptimal decision loses exactly $2t$ in expected arrival reward. The transition rows do not depend on the current state, so the best expected arrival reward at every step is $1/n + t$. For any history-dependent execution rule its normalized expected loss is therefore exactly

$$2ta \sum_{k \geq 0} \gamma^k \Pr_\sigma(B_\sigma(k)). \tag{222}$$

Since $a \sum_{k<T} \gamma^k \geq 1/2$, (221)–(222) give conditional bit-averaged loss at least $t/4$. This bound is uniform over the pair and its conditional private-stream law. Events with zero conditioning probability contribute nothing. Using (218) and nonnegativity of the conditional expected loss off E, the unconditional finite-prior mean is at least

$$\Pr(E)\frac{t}{4} \geq \frac{t}{4\binom{n}{2}} = \frac{1}{64n\binom{n}{2}\sqrt{H}}. \tag{223}$$

At least one fixed pair/bit instance attains this loss over the algorithm's own random stream and trajectory. The instance is not chosen after either is observed. This proves the first lower envelope in (195), completing the query-count law in (198).

The exact-table lower is compatible with zero loss at $\delta = 0$ once $n - 1$ queries are allowed. Its obstruction is the unobserved pair after acquisition, and its rate already accounts for learning during execution.

H RELATIVE RESIDUALS AT SMALL INVERSE MASS

The proof compares residuals relatively, even when their absolute scale shrinks. To illustrate this mechanism, we add a common reset of probability $\lambda = H^{-1/2}$ to the three-core construction. The inverse masses are $c_g = H^{-1}/d$ on the core and $c_f = H^{-1}$ elsewhere, where $d = 1 - \gamma(1 - \lambda)$, so $\max c \to 0$. The 72 conditions vary state count, horizon, initial pair, probe orientation and noise; half use a shared correlated perturbation of the complete critic table.

After two core queries, the remaining core label g must be preserved. Its exact relative gap over each label $f \in F$ is $(b_g - b_f)/b_g = 2\gamma(1 - \lambda)/3$. Across the 84 comparisons, the smallest b_g decreases from 0.2480 to 0.01844. The relative gap stays above 0.4687, while the relative score error $\|\widetilde{b} - b\|_\infty / b_g$ stays below 0.00237 (Figure 2). Every choice preserves g. The rule starting from the empty set avoids this configuration by querying labels in F earlier. Appendix I reports the full design, query traces, fitted models, and comparisons with prespecified omission and a rule that adapts only its final query.

Under the declared error budget, the acquired columns and fitted kernel also give the computable recovery bound in Corollary 4. In all 36 conditions with added correlated noise, both sequential variants have bounds below 0.086 and measured maximum transition-row ℓ_1 errors below 0.0109. These are post-hoc checks of the fitted kernels. Appendix C proves the certificate, and Appendix I.3 reports its values and comparisons across protocols.

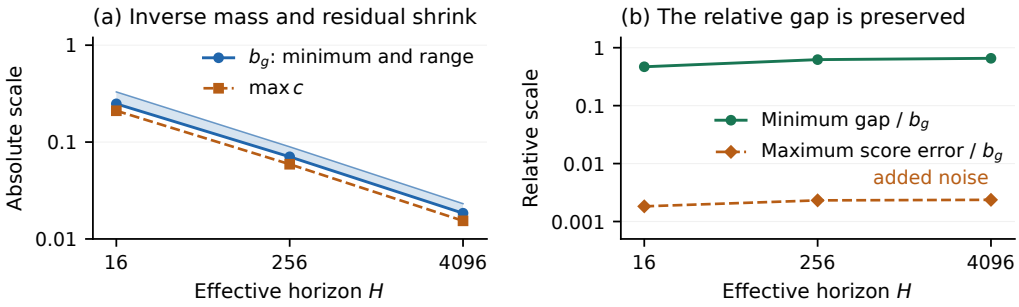

Figure 2: **The comparison survives a shrinking absolute scale.** Left: inverse mass and the range of core residuals over all critical comparisons. Right: minimum relative gap and maximum relative score error under added noise. Both axes are logarithmic; the shaded band spans the tested conditions. Both starts leave a label with maximum inverse mass unqueried in every case.

I FINITE CHECKS OF ACQUISITION AND THE RESIDUAL GAP

The study contains 72 conditions in which the maximum inverse mass shrinks with the horizon. Its design was fixed before the run, and all conditions, including zero losses, are retained. Losses are normalized gaps between the optimal return and that of each fitted stationary policy. Expected infinite-horizon discounted returns are computed by solving linear equations. The conditions share orientations and noise and are not independent replications. The study examines relative residual comparisons and stationary planning; the uniform retention guarantee follows from Appendix E.

I.1 A COMMON CONSTRUCTION AND EVALUATION PROTOCOL

Use a three-state core $R = \{0, 1, 2\}$, forced states $F = S \setminus R$, five actions, and $w = \mathrm{Unif}(R)$. At every core state, three actions have rows $\mathrm{Unif}(R \setminus \{g\})$, one for each $g \in R$. Two probe actions have rows

$$p_\theta = (1-\theta)w + \theta e_j^\top, \qquad \theta_\pm = 1/16 \pm t, \qquad t = 1/64, \tag{224}$$

where $j \in F$. The bit $\sigma \in \{0, 1\}$ exchanges the probe labels. Every action at a forced state has row w. Denote the resulting kernel by K_σ; its global transition-row diameter is $1/2$.

The sequential variant without a prescribed prefix repeatedly queries a minimum signed residual

$$\widetilde{b}_i = 1 - \widetilde{M}_{iI}(\widetilde{M}_{II})^{-1}\mathbf{1}, \qquad i \notin I, \tag{225}$$

starting from scores one and stopping after $n-1$ distinct queries. Exact ties in the computed scores use the smallest label among the tied unqueried candidates. The initialized variant first queries labels 0 and 1, then uses the same rule. Only acquired columns enter either computation. Full acquisition receives all n columns. Uniform nonadaptive omission is evaluated by fitting every one of the n possible omissions and averaging their losses exactly. The last-query heuristic first acquires $I_0 = \{0, \ldots, n-3\}$, computes a unit-ℓ_1 left-null vector of $[\mathbf{1}, \widetilde{M}_{I_0}]$ by singular-value decomposition, and queries the remaining label with larger absolute null coordinate. This is a particular member of the late-adaptive class.

All methods receive restrictions of the same complete noisy table and use the rowwise maximum-residual simplex fit (129), with HiGHS primal and dual feasibility tolerances 10^{-9}. After acquisition, reveal reward $r = e_j$, plan by policy iteration in $\widehat{P}$, and evaluate the returned policy in the true kernel at the relabeled image o of canonical core state 0:

$$L = H^{-1}\{J_P^*(o) - J_P^{\widehat{\pi}}(o)\}, \qquad J_P^{\widehat{\pi}} = P_{\widehat{\pi}}(I - \gamma P_{\widehat{\pi}})^{-1} r. \tag{226}$$

The selector receives neither the reward nor the root, true model, special-state designation, error budget, or diagnostic quantities.

I.2 SHRINKING INVERSE MASS: ALL 72 CONDITIONS

The Cartesian product is $n \in \{5, 8\}$, $H \in \{16, 256, 4096\}$, three label layouts, two bits, and two noise conditions. Set $j = n-1$ in canonical labels. The observed-to-canonical permutations are $[0, 1, 2, 3, 4, \ldots, n-1]$, $[3, 4, 0, 1, 2, 5, \ldots, n-1]$, and $[0, 3, 1, 2, 4, 5, \ldots, n-1]$. Thus the fixed two-label prefix contains two core states, two forced states, or one of each. The special state and root are relabeled by the same permutation.

A common reset makes inverse mass small. Put $a = H^{-1}$, $\gamma = 1-a$, $\lambda = H^{-1/2}$, $\nu = \mathrm{Unif}(S)$, and

$$P_\sigma = (1-\lambda)K_\sigma + \lambda \mathbf{1}\nu, \qquad \beta = \gamma(1-\lambda), \qquad d = 1-\beta = a + \gamma\lambda. \tag{227}$$

Its diameter is $(1-\lambda)/2$. For $T = \sum_{f \in F} e_f w$ and $B = I + \beta T$, the exact minimum-occupation matrix is

$$M = B + \frac{\gamma\lambda}{a}\mathbf{1}\nu B, \qquad c = M^{-1}\mathbf{1} = \frac{a}{d}B^{-1}\mathbf{1}, \qquad c_g = a/d\,(g \in R), \quad c_f = a\,(f \in F). \tag{228}$$

Indeed $T^2 = 0$, and the exclusion actions minimize occupation of a core goal. Hence $B = I + \beta \min K_\sigma B$; adding the columnwise constant in (228) gives the Bellman identity for P_σ. The inverse formula follows by a rank-one update. Both bits have the same M. Its maximum inverse mass tends to zero, whereas the forced-to-core mass ratio is d.

Complete noisy input. Compute $q_\sigma = P_\sigma M$ and $\bar{q}_\sigma = q_\sigma / H$. Start with their two-bit midpoint and replace column j by its exact value for the current bit. Call this table $Z_\sigma^{(0)}$. Since $(e_j - w)B = e_j - dw$, its maximum normalized error is

$$\delta_{\mathrm{base}} = \frac{(1-\lambda)td}{3H}. \tag{229}$$

Table 4: Shrinking absolute scales and preserved relative comparisons. Each horizon has 28 critical comparisons. The last column is the maximum of $\|\widetilde{b}-b\|_\infty/b_g$ over the 14 added-noise comparisons; midpoint-only errors are at most 6.03×10^{-15}.

H	$\max c$	Minimum b_g	Minimum gap/b_g	Maximum error/b_g
16	0.210526	0.248002	0.468750	0.001827
256	0.059041	0.070455	0.622559	0.002317
4096	0.015388	0.018438	0.656090	0.002367

Table 5: All 72 conditions. The uniform nonadaptive loss is the exact average over all omissions within a condition; the displayed mean and maximum then range over conditions. Positive means greater than 10^{-8}. The three zero-loss methods are shown explicitly.

Protocol	Positive cases	Mean loss	Maximum loss
Full acquisition	$0/72$	0	0
Sequential	$0/72$	0	0
Two-label prefix, then sequential	$0/72$	0	0
Uniform nonadaptive omission	$39/72$	0.002273	0.010349
Last-query heuristic	$13/72$	0.004856	0.051574

The second noise condition adds one complete correlated tensor

$$E_{sug} = \frac{0.002}{H}(0.6G_g + 0.25S_s + 0.15A_u), \qquad \delta = \delta_{\text{base}} + 0.002/H. \tag{230}$$

The entries of the three vectors are independent uniform draws from $[-1, 1]$, using NumPy's default generator with seed $20260910+100000n+100H+10h_{\text{lay}}$, where the layout index is $h_{\text{lay}} \in \{0, 1, 2\}$. The same tensor is used for both bits. Midpoint-only input uses $E = 0$ and budget δ_{base}. All noise is fixed before querying.

For each condition, acquired labels are sorted before fitting to keep LP constraint order fixed. Identical retained sets share a cached fit, giving 540 distinct fits. Negative probability entries of magnitude at most 10^{-9} may be clipped, and rows are normalized only if their sum discrepancy is at most 10^{-8}. Larger deviations are classified as numerical failures. Policy iteration starts at action zero and stops at maximum arrival-action improvement 10^{-9}, with at most 200 updates. Raw loss below -10^{-7} is classified as a numerical failure; negative roundoff is otherwise clipped to zero. Positive counts use threshold 10^{-8}. No fit, planning failure, or query fallback occurred.

Observed relative residual gaps. A critical comparison is a sequential decision with one core label g and at least one forced label still unqueried. Fixed prefix decisions are excluded. There are 28 comparisons at each horizon, all in the two-core prefix layout: 14 with midpoint-only input and 14 with added noise. The empty-prefix sequential rule has zero such comparisons, since it queries forced labels while at least two core labels remain.

The gap has a direct identity. Let $A = B^{-1} = I - \beta T$, $z = A\mathbf{1}$, and $\rho = \gamma\lambda/d$. Then $C = M^{-1} = A - \rho z\nu$. If the remaining set has one core label and f forced labels, $C_{UU}b_U = c_U$ implies, with their common forced residual b_F,

$$b_g - \frac{\rho}{n}(b_g + fb_F) = a/d, \qquad b_F - \frac{\beta}{3}b_g - \frac{\gamma\lambda}{n}(b_g + fb_F) = a.$$

Subtracting d times the first equation from the second gives

$$b_F = (d + \beta/3)b_g, \qquad \frac{b_g - b_F}{b_g} = \frac{2\beta}{3}. \tag{231}$$

All 84 comparisons agree with this identity within 4.89×10^{-15}. Table 4 gives their extrema; no critical comparison queries the core label.

Both sequential variants leave a core label with maximum inverse mass unqueried in every condition. Thus the experiment reaches small absolute inverse mass while preserving an omission with the

largest inverse mass. Midpoint-only features equal M in exact arithmetic, since the altered probes never minimize the affected core goals.

Reproducible evidence and scope. The design and all outcomes are available in `evidence/degenerate/`. The study includes 72 complete condition files with 8,784 arrays, including inputs, query traces, fitted kernels, policies, and values. The recorded diagnostics verify input budgets, Bellman and inverse identities, stochasticity, query decisions, policy-value equations, and uniform averaging. The largest policy-value equation residual is 1.14×10^{-13}, and recomputed loss arithmetic agrees exactly. Figure 2 uses these complete outputs.

These constructed-family checks evaluate stationary planning with a fixed probe amplitude $t = 1/64$. The history-dependent minimax law uses its own horizon-dependent testing amplitude and is established analytically; no online learner or critic-training procedure is evaluated here.

I.3 Certificate values on the cross-scale checks

We evaluate Corollary 4 on all 72 conditions without changing acquisition, fitting, or planning. The supplement records the reporting tolerances and the previously computed policy losses and aggregate recovery errors. Table 6 reports all 36 conditions with added correlated noise at tolerance 0.1. For full acquisition, we keep its fitted kernel and take the smallest bound over the n subsets of size $n - 1$. This uses columns already received and entails no additional fit. For uniform omission, we average the scalar bounds over omissions, obtaining a bound on expected maximum row error.

Table 6: Recovery errors and observed bounds in all 36 added-noise conditions. Error means maximum transition-row ℓ_1 error. For uniform omission, both error and bound are averaged over omissions within each condition before summarizing conditions. The last column counts conditions whose bound is at most 0.1 before rounding.

Protocol	Max. error	Mean bound	Max. bound	≤ 0.1
Full acquisition	0.008499	0.064994	0.105866	$31/36$
Sequential	0.010794	0.054817	0.085772	$36/36$
Two-label prefix, then sequential	0.010870	0.054827	0.085882	$36/36$
Uniform omission	0.219279	0.445029	1.194685	$6/36$
Last-query heuristic	0.332103	0.550004	1.868095	$12/36$

Both sequential variants have bounds below 0.086 in every added-noise condition; half also meet 0.05. Their median bound-to-error ratio is about 7.68, so these bounds remain conservative. In all 36 midpoint-only conditions, every protocol meets 0.1, and full acquisition and both sequential variants meet 0.05. The supplied records include every condition and reporting tolerance.

Bound size does not rank actual recovery error. Full acquisition has smaller error than sequential acquisition in 32 of the 36 added-noise conditions, although its bound is larger in all 36. Their mean budget contributions are both about 0.054724; the mean residual contributions are 0.010270 and 0.000093, respectively. The fits use different sets of constraints, so their observed residuals differ. The smaller bound does not establish an information advantage from omitting a column.

We evaluate the inverse, residual and tolerance comparisons in exact rational arithmetic from the stored inputs. Normalizing the stored fit rows changes them by at most 2.90×10^{-16} in ℓ_1 norm. The certificate is conditional on the supplied joint noise budget; comparison with the stored reference kernel remains a numerical check. The code, complete results and an independent arithmetic verification are provided in `supplement/posthoc_recovery/`.

J Proof of the minimum-occupation structure

The proof uses classical inverse-positive matrix ideas while retaining the same controlled action family through state elimination. The closure step below permits different actions to attain the different coordinate minima.

Lemma 4 (A controlled lower envelope)**.** *Let D be invertible with $B = D^{-1}$ having nonpositive off-diagonal entries and $B\mathbf{1} > 0$. Let $e_u \geq 0$ and $e_u\mathbf{1} \leq \gamma$ for a finite nonempty family of rows, and set $b_j = \min_u (e_u D)_j$. Then $w = bB \geq 0$ and $w\mathbf{1} \leq \gamma$.*

Proof. For each j, choose u_j attaining the minimum. The diagonal term vanishes in the identity

$$w_j - (e_{u_j})_j = \sum_{k \neq j} [b_k - (e_{u_j} D)_k] B_{kj} \geq 0.$$

For any u, $b \leq e_u D$ and $B\mathbf{1} > 0$ give $w\mathbf{1} = bB\mathbf{1} \leq e_u DB\mathbf{1} = e_u\mathbf{1} \leq \gamma$. □

Proof of Theorem 2. Each column solves a γ-contraction and is nonnegative. First use induction on the state count to prove invertibility, nonpositive off-diagonal entries of the inverse, and strictly positive inverse row sums. The one-state case is immediate.

Fix any state i, let $R = S \setminus \{i\}$, and partition $M = \left(\begin{smallmatrix} a & b \\ c & D \end{smallmatrix}\right)$. For goals in R, eliminate state i by defining

$$e_u = \frac{K_{i,u,R}}{1 - K_{i,u,i}}, \qquad b_g = \min_u e_u D_{\cdot g}, \qquad e_u\mathbf{1} \leq \gamma. \tag{232}$$

Indeed, the scalar Bellman equation gives $b_g \leq K_{i,u,i} b_g + K_{i,u,R} D_{\cdot g}$ for every u, with equality for a minimizing action. Dividing by $1 - K_{i,u,i} > 0$ gives exactly the minimum in (232). Substitution into the remaining Bellman equations gives action rows $K'_{s,(v,u)} = K_{s,v,R} + K_{s,v,i} e_u$, again with row sum at most γ. Thus D is a minimum-occupation matrix of the same kind on one fewer state. The substitution is an equality of Bellman equations; it does not require any extra dependence of the original policy on the preceding state.

By induction, $B = D^{-1}$ has nonpositive off-diagonal entries and $B\mathbf{1} > 0$. Lemma 4 gives $w = bB \geq 0$ and $w\mathbf{1} \leq \gamma$, even when different coordinates attain their minima at different actions. For column i, $a \geq 1$, and a maximum at another state would imply $m \leq \gamma m$ with $m \geq 1$, which is impossible. Hence $0 \leq c \leq \gamma a\mathbf{1}$. Its Schur complement satisfies $T = a - wc \geq a(1 - \gamma^2) > 0$. The ith row of M^{-1} is $(1, -w)/T$, with the required signs and positive row sum. Since i was arbitrary, induction closes.

We can now derive the sharp quantitative bound without assuming these signs in advance. Choose u optimal at state i for goal i, and let $z = K_{i,u} M - M_{i,\cdot} + e_i$, where e_i is a coordinate row. Bellman optimality gives $z \geq 0$ and $z_i = 0$. Hence $(zC)_i \leq 0$, while $zC = K_{i,u} - e_i + C_{i,\cdot}$, proving $C_{ii} \leq 1 - K_{i,u,i} \leq 1$. Moreover $zC\mathbf{1} \geq 0$ yields $(C\mathbf{1})_i \geq 1 - K_{i,u}\mathbf{1} \geq 1 - \gamma$. The absolute row sum is therefore $2C_{ii} - (C\mathbf{1})_i \leq 1 + \gamma$. For sharpness, the single-action chain $1 \to 2$, $2 \to 2$ has $C = I - \gamma P$, whose first row is $(1, -\gamma)$. □

K Common transition mass and fixed-omission precision

The main comparison uses only a global row-diameter bound. A transition component shared by every row further reduces the cost of a prespecified omission. Define

$$\kappa(P) = \max_{s,u,t,v} \mathrm{TV}(P_{su}, P_{tv}), \qquad \mu(P) = \sum_x \min_{s,u} P(x \mid s, u). \tag{233}$$

Thus $\mu(P) \geq \lambda$ means that $P = (1 - \lambda)K + \lambda\mathbf{1}\nu$ for a stochastic kernel K and a probability row ν. Pairwise overlap need not provide common mass. The following refinement concerns the complete table and one prespecified omission, under the same evaluated-interaction convention.

Theorem 7 (Critic precision under evaluated interaction)**.** *Restrict the execution risks to kernels with $\kappa(P) \leq 1/2$ and $\mu(P) \geq \lambda$. Suppose $n \geq 4$, $A \geq 5$, $H \geq 2$ and $0 \leq \lambda \leq 1/2$. Set $d = 1 - \gamma + \gamma\lambda$. For every $\delta \geq 0$,*

$$\begin{aligned} \mathcal{D}^{\mathrm{exec}}_{\mathrm{full},\lambda} &= \widetilde{\Theta}_{n,A}\Big(\min\{\delta H, H^{-1/2}\}\Big), \\ \mathcal{D}^{\mathrm{exec}}_{\mathrm{del},\lambda} &= \widetilde{\Theta}_{n,A}\Big(\min\{\delta H/d, H^{-1/2}\}\Big). \end{aligned} \tag{234}$$

Precisely, each displayed minimum is a lower bound after multiplication by $1/640$*, and an upper bound after multiplication by* $C_{n,A}\sqrt{\log(eH)}$*. The constants are independent of* δ, H, λ*; the explicit branch bounds appear in Appendix L. The lower bounds allow randomized, history-dependent execution. Their models have* $\mu(P) = \lambda$*,* $\kappa(P) = (1-\lambda)/2$*, and every transition entry is at least* λ/n*.*

At $\lambda = 0$, this gives the complete-library and fixed-omission rates used in Theorem 6 at $\kappa = 1/2$. Fixed positive common mass removes the additional horizon factor. Common mass of order H^{-1} leaves that factor even when every transition entry is positive. The proof, including the bounded-bias online branch, follows in Appendix L.

L Critic precision and information learned during execution

We prove Theorem 7. Write $a = 1-\gamma$, $H = a^{-1}$, and

$$\begin{aligned} &\kappa(P) = \max_{x,u,y,v} \mathrm{TV}(P_{xu}, P_{yv}), \qquad \mu(P) = \sum_z \min_{x,u} P(z \mid x, u), \\ &\mathcal{P}_{n,A}(\lambda) = \{P : \kappa(P) \le 1/2,\ \mu(P) \ge \lambda\}. \end{aligned} \tag{235}$$

The risks are the execution risks in (282), with the kernel supremum restricted to (235). The observation is the complete normalized table $\bar{q} = aq$, or its restriction after a prespecified singleton j is deleted, with entrywise error at most δ. The instance (P, Z, r, o) is fixed before the algorithm's private randomness. Execution may use the entire evaluated trajectory, but no preliminary environment interaction or reset labels are supplied. Current-state and arrival values retain the convention $W_P^\pi = r + \gamma J_P^\pi$ from (281).

The theorem assumes $n \ge 4$, $A \ge 5$, $H \ge 2$, $0 \le \lambda \le 1/2$, and arbitrary $\delta \ge 0$. Put

$$d_\lambda = a + \gamma\lambda. \tag{236}$$

The upper proofs hold for arbitrary finite nonempty action sets; the dimension and action requirements come from the lower construction. The discount, state and action labels, and downstream reward are known. At each stated error budget the minimax rule may select an upper-bound branch using δ and the promised class parameter. This does not assert one rule that adapts to every unknown noise level.

For $m = \sum_x |\mathcal{A}_x|$ state-action pairs, define

$$\begin{aligned} B_{n,m}(H) = 8\sqrt{\frac{m}{H}\left[n\log 2 + \log(4mH^2) + 2\log(2H)\right]} \\ + \frac{2m\{1 + \log_2(H+1)\} + 4}{H} + \frac{3}{2H^2}. \end{aligned} \tag{237}$$

All logarithms except $\log_2$ are natural. We show below that an execution rule can ignore the critic table and attain loss at most $B_{n,m}(H)$. For fixed dimensions this is $O_{n,m}(\sqrt{\log(eH)/H})$, uniformly for $H \ge 2$.

L.1 Coordinates under an unknown common minorization

The quantity $\mu(P)$ is the classical common one-step minorization mass (Lee et al., 2026). If $\mu(P) > 0$, the row defined by $\nu_z = \min_{x,u} P(z \mid x, u)/\mu(P)$ is a probability distribution. Whenever $0 \le \lambda \le \mu(P)$ and $\lambda < 1$, subtracting $\lambda\nu$ from every row gives a stochastic kernel K such that

$$P = (1-\lambda)K + \lambda \mathbf{1}\nu. \tag{238}$$

For $\lambda = 0$, take $K = P$ and choose any ν. Thus every member of (235) has such a decomposition with the promised mass λ. Neither the residual kernel nor its witness ν is an observed input. Different true models may have different witnesses.

Fix one decomposition and let $\beta = \gamma(1-\lambda)$, so that $d_\lambda = 1 - \beta$, and define

$$T = I + \frac{\gamma\lambda}{a}\mathbf{1}\nu, \qquad T\mathbf{1} = \frac{d_\lambda}{a}\mathbf{1}. \tag{239}$$

If v is an optimized current-state cost value for (K, β), then its value for (P, γ) is

$$v_P = v + \frac{\gamma\lambda}{a}(\nu v)\mathbf{1} = Tv. \tag{240}$$

Indeed, setting $h = (\gamma\lambda/a)\nu v$, Bellman substitution gives

$$c + \gamma \min_u P_u(v + h\mathbf{1}) = v + (\gamma\lambda\nu v + \gamma h)\mathbf{1} = v + h\mathbf{1}.$$

The constant does not change any action comparison, and discounted uniqueness proves (240). The same calculation works for maximized rewards and for any fixed stationary policy. Consequently the *same* operator transforms all optimized singleton columns and every fixed-policy Green matrix:

$$M_P = TM_{K,\beta}, \qquad N_{P,\pi} = TN_{K,\beta,\pi}. \tag{241}$$

No common policy across the singleton tasks is assumed.

Let $C_P = M_P^{-1}$, $c_P = C_P\mathbf{1}$, and define C_β, c_β in the latent model. Lemma 5 at discount β gives $d_\lambda\mathbf{1} \le c_\beta \le \mathbf{1}$. Equations (239)–(241) therefore imply

$$c_P = \frac{a}{d_\lambda}c_\beta, \qquad a\mathbf{1} \le c_P \le \frac{a}{d_\lambda}\mathbf{1}, \qquad C_P N_{P,\pi} = C_\beta N_{K,\beta,\pi} \ge 0. \tag{242}$$

For any stationary policy and reward $0 \le r \le \mathbf{1}$, the same lemma gives $0 \le C_P W_P^\pi \le Hc_P$. Its deleted-coordinate construction, with $J = S \setminus \{j\}$, now yields

$$W_P^\pi = k\mathbf{1} + M_{P,J}\alpha, \qquad \|\alpha\|_1 \le H\sum_{i\ne j}(c_P)_i \le \frac{n-1}{d_\lambda}. \tag{243}$$

Explicitly, for $b = C_P W_P^\pi$ use $k = b_j/(c_P)_j$ and $\alpha_i = b_i - k(c_P)_i$. For the full library, centering b at $Hc_P/2$ gives coefficient norm at most $n/(2d_\lambda)$. These are true-model coordinates used only in the analysis. They are uniform over all stationary policies, including a policy chosen using a fitted model, and over all bounded rewards.

L.2 Upper bounds using the observed noisy table

Use the same supremum-norm simplex fit as (286). For retained indices $\mathcal{I}$, set

$$\widetilde{q} = HZ, \qquad \widetilde{M}_{xg} = \mathbf{1}\{x = g\} + \gamma \min_u \widetilde{q}_g(x, u).$$

$$\widehat{p}_{xu} \in \arg\min_{z\in\Delta_n} \|z\widetilde{M} - \widetilde{q}(x, u, \cdot)\|_\infty.$$

Here γ and H are the MDP's discount and normalization. The decoder is not given a transformed table for (K, β). It needs none of λ, ν, K, β, or δ; fix any deterministic minimizer rule. The true row has fitting residual at most $(1 + \gamma)H\delta$, and the fitted residual is no larger. As $\|\widetilde{M} - M_{P,\mathcal{I}}\|_{\max} \le \gamma H\delta$, the direct calculation in (287) gives

$$\|(\widehat{p} - p)M_{P,\mathcal{I}}\|_\infty \le 2(1 + \gamma)H\delta. \tag{244}$$

The fit is over the entire stochastic simplex; no unknown minorization constraint is imposed on the fitted kernel.

After deleting j, row differences annihilate constants, so (243) and (244) give, simultaneously for all true-model policy values,

$$|(\widehat{p}_{xu} - p_{xu})W_P^\pi| \le \frac{2(n-1)(1+\gamma)}{d_\lambda}\delta H.$$

The arrival comparison (289) bounds normalized evaluation error by this quantity. Comparing a true optimal policy with a fitted-model optimal policy, as in Appendix N, doubles it. Applying this to each true kernel proves

$$\mathcal{D}^{\mathrm{exec}}_{\mathrm{del},\lambda}(\delta;\gamma) \le \frac{4(n-1)(1+\gamma)}{d_\lambda}\delta H \le \frac{8(n-1)}{d_\lambda}\delta H. \tag{245}$$

This reasoning is instancewise in the unknown witness ν and then uniform over the class. The fitted kernel need only be stochastic; it need not satisfy either promise. The fit is independent of the reward, so the same fitted model supports rewards disclosed afterward.

For the full-library upper, use the row-diameter promise instead. The Dobrushin span contraction is classical; see Gaubert & Qu (2015). Every stationary randomized policy has row diameter at most $1/2$, since its rows are convex combinations of the legal transition rows. Hence

$$\operatorname{span}(W_P^\pi) \le \operatorname{span}(r) + \frac{\gamma}{2}\operatorname{span}(W_P^\pi), \qquad \operatorname{span}(W_P^\pi) \le \frac{1}{1-\gamma/2}.$$

Center W_P^π at the midpoint k of its minimum and maximum. Theorem 2 gives

$$W_P^\pi = k\mathbf{1} + M_P\alpha, \qquad \alpha = C_P(W_P^\pi - k\mathbf{1}), \qquad \|\alpha\|_1 \le \frac{n(1+\gamma)}{2(1-\gamma/2)}.$$

Apply (244) and the same evaluation and planning comparison to obtain

$$\mathcal{D}^{\mathrm{exec}}_{\mathrm{full},\lambda}(\delta;\gamma) \le \frac{2n(1+\gamma)^2}{1-\gamma/2}\delta H \le 16n\delta H. \tag{246}$$

Both upper arguments use the stated arrival interface, so the arrival-greedy convention in Appendix N also handles $\gamma = 0$. This endpoint is not needed for the lower bound. Neither argument restricts the noise level.

There is also a common cap. In a fixed environment, (238) implies that every conditional expected arrival reward is $\lambda\nu r + (1-\lambda)z$ for some $z \in [0,1]$. Every normalized return, including one from a history-dependent policy, lies in an interval of width $1-\lambda$. Thus every execution loss is at most $1-\lambda$, without knowledge of ν.

L.3 An online upper bound on the evaluated trajectory

Bounded-bias optimism is a classical approach to online control; SCAL provides one established example (Fruit et al., 2018). We give a self-contained proof for the present class to specify the discounted comparator and constants. The construction uses exact global optimization over a confidence set. It proves existence of an execution rule, without claiming an efficient planning oracle. It observes only transitions on the trajectory whose return is evaluated and never reads the initial critic table.

For a candidate kernel Q with $\kappa(Q) \le 1/2$, consider the current-reward operator

$$(F_Q h)(x) = r(x) + \max_u Q_{xu} h.$$

For any two vectors h, h', differences between coordinates of $F_Q h - F_Q h'$ are bounded by differences of expectations of $h - h'$ under two rows of Q. Consequently

$$\operatorname{span}(F_Q h - F_Q h') \le \tfrac{1}{2}\operatorname{span}(h - h').$$

Fixing one coordinate at zero gives a contraction modulo constants. Its fixed point supplies an average gain g_Q and bias h_Q with

$$g_Q\mathbf{1} + h_Q = r + \max_u Q_u h_Q, \qquad \operatorname{span}(h_Q) \le 2. \tag{247}$$

The span bound follows by comparing with $F_Q 0 = r$. Shift h_Q so that $0 \le h_Q \le 2$. A stationary maximizing policy in (247) has gain g_Q from every root: the bias terms in its expected finite-prefix rewards telescope. The inequality for other actions proves average optimality even against history-dependent policies. Multiplying the equality by an invariant law of the maximizing stationary kernel gives $g_Q \in [0,1]$. No positive stationary mass for every state is needed. Uniform contraction also makes the normalized fixed point and gain continuous in Q.

Set $\zeta = 1/(2H^2)$. For a state-action pair, let $\widehat{P}_N$ be its empirical next-state row after N visits. At $N = 0$, use the fixed uniform row. Define

$$\begin{aligned} b_0 &= 2, \\ b_N &= \min\left\{2, \sqrt{\frac{2}{N}\left[n\log 2 + \log\frac{2mN^2}{\zeta}\right]}\right\}, \qquad N \ge 1. \end{aligned} \tag{248}$$

For each pair, its successive next states can be represented by an independent sample sequence from the true row, revealed only when that pair is visited. This representation gives no extra samples to the algorithm. The identity $\|v\|_1 = \max_{\sigma\in\{-1,1\}^n} \sigma^\top v$ and the bounded-variable exponential inequality give

$$\Pr\{\|\widehat{P}_N - P_{xu}\|_1 > \varepsilon\} \le 2^n \exp(-N\varepsilon^2/2).$$

Union bounding over all pairs and positive sample counts, using $\sum_{N\ge1} N^{-2} \le 2$, proves that

$$\begin{aligned} \mathcal{E} &= \{\|\widehat{P}_N(x,u) - P_{xu}\|_1 \le b_N \text{ for every pair and every } N \ge 1\}, \\ \Pr(\mathcal{E}^c) &\le \zeta. \end{aligned} \tag{249}$$

The zero-count radius includes every stochastic row. In particular this is one all-time event, independent of episode lengths.

At an episode start, with current counts $N(x,u)$, form the set of stochastic kernels Q satisfying $\kappa(Q) \le 1/2$ and

$$\|Q_{xu} - \widehat{P}_{N(x,u)}(x,u)\|_1 \le b_{N(x,u)} \qquad \text{for every } (x,u). \tag{250}$$

If the set is nonempty, choose a model maximizing g_Q. Compactness and gain continuity ensure a maximizer. If the set is empty, choose the kernel whose every row is uniform. All possible histories have countable data tables, so ties can be resolved deterministically. On $\mathcal{E}$ the set contains the true kernel and $g_Q \ge g_P$. On every path, including failure paths, the chosen bias satisfies (247).

Execute a stationary maximizing policy for the chosen model until some pair has received $\max\{1, N(x,u)\}$ visits during this episode, where N is its count at the episode start. End the episode immediately and recompute. Write N_e for the starting counts of episode e. This visit-doubling rule does not require a state to be visited at all.

Write $g_* = g_P$ and $R_T = Tg_* - \sum_{t=0}^{T-1} r(S_t)$ for current-reward regret. At step t of episode e, with model Q_e, gain g_e, and bias h_e, the chosen action gives

$$\begin{aligned} g_* - r(S_t) = (g_* - g_e) + (Q_e - P)_{S_tA_t} h_e \\ + \{P_{S_tA_t} h_e - h_e(S_{t+1})\} + h_e(S_{t+1}) - h_e(S_t). \end{aligned} \tag{251}$$

On $\mathcal{E}$, the first term is nonpositive and the row-error term is at most $2b_N$: the rows share a radius-b_N confidence ball and $\operatorname{span}(h_e)/2 \le 1$. On its complement those two terms together are at most three. On every path they are therefore at most $2b_N + 3\mathbf{1}_{\mathcal{E}^c}$. The term in braces is a martingale difference with unconditional mean zero. The remaining bias terms telescope within each episode and contribute at most two per episode. This does not condition the martingale on the confidence event.

For $T \ge 1$, put

$$L_T = n\log 2 + \log(4mH^2) + 2\log(T+1).$$

If a pair has count q just before a step and count N at that episode's start, the stopping rule implies

$$\frac{1}{\sqrt{\max\{1,N\}}} \le \frac{\sqrt{2}}{\sqrt{q+1}}.$$

Also $b_N \le \sqrt{2L_T}/\sqrt{\max\{1,N\}}$, including $N = 0$. Summing $(q+1)^{-1/2}$ over each pair's visits and then using Cauchy–Schwarz across pairs yields

$$\sum_{t<T} b_{N_e(S_t,A_t)} \le 4\sqrt{mTL_T}.$$

A pair can end an episode on its first visit and then only after its count has at least doubled. Thus at most $1 + m\{1 + \log_2(T+1)\}$ episodes have begun by time T. Taking unconditional expectations in (251) proves

$$\mathbb{E}R_T \le 8\sqrt{mTL_T} + 2\left[1 + m\{1 + \log_2(T+1)\}\right] + 3\zeta T. \tag{252}$$

Rarely visited states enter through their actual visit counts; there is no lower bound on those counts.

To pass to discounted rewards, let K be an independent geometric index with $\Pr(K = k) = a\gamma^{k-1}$ for $k \geq 1$. Abel summation gives

$$g_* - a\mathbb{E}\sum_{t\geq 0}\gamma^t r(S_t) = a^2\sum_{T\geq 1}\gamma^{T-1}\mathbb{E}R_T = a\mathbb{E}R_K. \tag{253}$$

The whole infinite tail is included, and absolute convergence follows from $|R_T| \leq T$. Here $\mathbb{E}K = H$ and $\mathbb{E}K^2 = 2H^2 - H$. Under the size-biased law $k\Pr(K = k)/H$, the mean of $K + 1$ is $2H$. Jensen's inequality therefore gives

$$\begin{aligned}\mathbb{E}[K\log(K+1)] &\leq H\log(2H),\\ \mathbb{E}\sqrt{KL_K} &\leq \sqrt{H\{n\log 2 + \log(4mH^2) + 2\log(2H)\}},\\ \mathbb{E}\log_2(K+1) &\leq \log_2(H+1).\end{aligned} \tag{254}$$

The endpoint correction uses the arrival convention explicitly. For the true current-reward bias, set $b = h_* - r$. Then

$$g_*\mathbf{1} + b = \max_u P_u(r + b), \qquad \operatorname{span}(b) = \operatorname{span}(\max_u P_u h_*) \leq 1. \tag{255}$$

Shift b to $[0, 1]$. The action inequality in (255), summed with discount weights, gives $aJ_P^*(o) \leq g_* + a$. On every reward path,

$$\begin{aligned}a\sum_{t\geq 0}\gamma^t r(S_t) - a\sum_{t\geq 0}\gamma^t r(S_{t+1})&\\ = ar(S_0) - a^2\sum_{t\geq 1}\gamma^{t-1}r(S_t) &\in [-a, a].\end{aligned}$$

Consequently the algorithm's expected execution loss satisfies

$$a[J_P^*(o) - J_P^{\Pi}(o)] \leq a\mathbb{E}R_K + 2/H \leq B_{n,m}(H). \tag{256}$$

The last inequality is exactly (252)–(254), with $3\zeta = 3/(2H^2)$. The policy is deterministic conditional on its observed trajectory; J_P^{Π} averages that trajectory. The bound holds for each fixed reward, root, and input table, and requires no information about ν or positive stationary masses. It is an expected-return bound, not a high-probability realized-loss statement.

For later use, this explicit budget admits the following simpler uniform envelope. For $H \geq 2$, put $x = \log(eH)/H \leq 1$ and $Q_0 = n\log 2 + \log(16m)$. The square-root bracket in (237) is $Q_0 + 4\log H$, at most $(Q_0 + 4)\log(eH)$. Since $H + 1 \leq 2H$, its other terms are at most $(4m + 2m/\log 2 + 11/2)x$. Using $x \leq \sqrt{x}$ proves

$$B_{n,m}(H) \leq C_{n,m}\sqrt{\log(eH)/H} \qquad (H \geq 2) \tag{257}$$

for a finite constant depending only on the fixed state and action counts.

Corollary 5 (Online control under strict contraction)**.** *Fix $0 < \kappa < 1$, and put*

$$w_\kappa = \frac{1}{1-\kappa}, \qquad s_\kappa = \max\left\{1, \frac{w_\kappa}{2}\right\}, \qquad B_{n,m,\kappa}(H) = s_\kappa B_{n,m}(H). \tag{258}$$

For every kernel with row diameter at most κ, an execution rule that ignores the critic table has normalized loss at most $B_{n,m,\kappa}(H)$. Consequently,

$$B_{n,m,\kappa}(H) \leq C_{n,m,\kappa}\sqrt{\log(eH)/H} \qquad (H \geq 2)$$

for a finite constant depending only on n, m, κ.

Proof. Repeat the preceding construction with the confidence set restricted by $\kappa(Q) \leq \kappa$. The average-reward Bellman operator contracts spans by κ, so every optimistic bias has span at most w_κ. In the regret decomposition, the row-error, episode-boundary, and failure contributions are therefore bounded by $w_\kappa b_N$, w_κ, and $1 + w_\kappa$, respectively. Each is at most s_κ times the corresponding coefficient $2b_N$, 2, or 3 in the proof above.

For the arrival-reward endpoint, the shifted bias $b = h_* - r$ has span at most $\kappa w_\kappa = w_\kappa - 1$. Together with the one-step timing correction, this contributes at most w_κ/H, which is again at most s_κ times the preceding $2/H$ term. Multiplying the whole bound by s_κ proves the claim; the simpler envelope follows from (257). □

L.4 A STRICTLY POSITIVE LOWER FAMILY IN EVERY STATED DIMENSION

Let $s = 1 - \lambda$, $\beta = \gamma s$, and recall $d_\lambda = 1 - \beta$. Choose the prespecified omitted state j and a three-state core $R = \{x_1, x_2, x_3\} \subset S \setminus \{j\}$. Let w be uniform on R, as a row in $\mathbb{R}^n$. At j, every action of a base kernel has row w. At *every* non-j state, including all states outside the core, provide the same five labeled actions: the three rows

$$v_g = \mathrm{Unif}(R \setminus \{g\}), \quad g \in R,$$

and probes A and B with rows

$$\begin{aligned} &p_\theta = \theta e_j^\top + (1-\theta)w, \\ &\theta_+ = 1/8 + t, \quad \theta_- = 1/8 - t, \qquad 0 < t \le 1/32. \end{aligned} \tag{259}$$

Kernel K_0 assigns θ_+ to A and θ_- to B; K_1 exchanges them. If $A > 5$, duplicate v_{x_1} at non-j states and w at j. Finally set

$$\nu = \mathrm{Unif}(S), \qquad P_\sigma = sK_\sigma + \lambda \mathbf{1}\nu, \quad \sigma \in \{0,1\}. \tag{260}$$

All rows are stochastic. Distinct v_g have TV distance $1/2$, while

$$\mathrm{TV}(v_g, w) = 1/3, \quad \mathrm{TV}(p_\theta, w) = \theta, \quad \mathrm{TV}(p_\theta, v_g) = \frac{1+2\theta}{3} \le \frac{7}{16}, \quad \mathrm{TV}(p_\theta, p_{\theta'}) = |\theta - \theta'|.$$

Thus $\kappa(P_\sigma) = s/2$. At each coordinate some base row has zero mass: use v_g at a core coordinate, w at j, and any base row at a coordinate outside $R \cup \{j\}$. Hence every coordinatewise row minimum of P_σ equals λ/n. It follows that

$$\mu(P_\sigma) = \lambda, \qquad \kappa(P_\sigma) = \frac{1-\lambda}{2}, \qquad P_\sigma(z \mid x, u) \ge \frac{\lambda}{n}. \tag{261}$$

In particular the lower models are strictly positive when $\lambda > 0$. The extra states use the specified action menu; they are not absorbing padding states, and the reset can visit them.

At discount β, the actual base minimum-occupation matrix is

$$M_K = I + \beta e_j w. \tag{262}$$

For a core goal g, every non-j state can choose v_g and avoid that goal thereafter; the forced row at j contributes $\beta/3$. For goal j, choose base actions that never return to j. A goal outside the core and j is never entered by a base transition. These observations verify all Bellman minima for (262): other continuation costs are nonnegative. Discounted uniqueness proves the matrix formula, which is independent of both probe parameters.

Applying P to (240) gives the *complete* arrival-critic identity

$$q_{P,\gamma,g}(x,u) = s q_{K,\beta,g}(x,u) + \frac{\lambda}{a}\nu(M_K)_g. \tag{263}$$

The coefficient of the constant is λ/a, not $\gamma\lambda/a$. Although target dependent, this constant agrees between the two models because M_K does. At a base probe,

$$q_{K,j}(p_\theta) = \theta, \qquad q_{K,g}(p_\theta) = \frac{1 - d_\lambda \theta}{3} \ (g \in R), \qquad q_{K,g}(p_\theta) = 0 \ (g \notin R \cup \{j\}).$$

Every nonprobe row agrees between models. Equation (263) therefore proves the exact complete-table distances

$$\|\bar{q}_0 - \bar{q}_1\|_{\max} = 2ast, \qquad \|\bar{q}_{0,-j} - \bar{q}_{1,-j}\|_{\max} = \frac{2asd_\lambda t}{3}. \tag{264}$$

There is no unexamined critic row at an extra state or action. The full midpoint is itself the exact table of the model with both probe parameters equal to $1/8$, since M_K is parameter independent and every critic above is affine in its probe parameter. Its restriction is the deleted midpoint. These common inputs will be fixed before any private randomization.

L.5 A TESTING LOWER BOUND THAT ALLOWS LEARNING DURING EXECUTION

Give both models the same reward $r = e_j$ and initial state x_1. Every non-j state has the same optimal base current-state value

$$x_\beta = \frac{\beta\theta_+}{d_\lambda(1+\beta\theta_+)}, \qquad W^*_{K,j} = 1 + \beta x_\beta. \tag{265}$$

Indeed, using the larger probe gives $x_\beta = \beta[\theta_+(1+\beta x_\beta) + (1-\theta_+)x_\beta]$. The resulting difference $W^*_{K,j} - x_\beta$ is positive, verifying that the larger probe dominates every other action outside j. The reset transformation adds a common constant to these values, so

$$W^*_{P,j} - W^*_{P,x} = \frac{1}{D} \quad (x \neq j), \qquad D = 1 + \beta\theta_+ \leq \frac{5}{4}. \tag{266}$$

The optimal value vector is the same in both models; the optimal probe label is exchanged. Define the raw arrival-action disadvantage

$$G_\sigma(x,u) = J^*_\sigma(x) - P_\sigma(\cdot \mid x,u)W^*_\sigma.$$

All actions at j have zero disadvantage. At a non-j state, any action other than model σ's larger probe has

$$G_\sigma(x,u) \geq \frac{2st}{D} \geq \frac{8st}{5}. \tag{267}$$

The smaller probe has equality in the first bound. A base action has probe parameter zero and at least this disadvantage, because $\theta_+ \geq 2t$.

Fix any randomized, history-dependent execution rule that receives a common admissible table. Suppose

$$t \leq \frac{1}{64\sqrt{H}}, \qquad T_0 = \lceil H \rceil \leq 2H. \tag{268}$$

We use the adaptive KL chain rule and Pinsker inequality (Lattimore & Szepesvári, 2020), keeping their constants explicit. For a base probe, the transition divergence is a Bernoulli divergence at parameters θ_+ and θ_-. In either direction,

$$\mathrm{kl}(p,q) \leq \frac{(p-q)^2}{q(1-q)} \leq 64t^2, \qquad p,q \in \{\theta_+,\theta_-\},$$

since $q \in [3/32, 5/32]$ and $q(1-q) \geq 81/1024$. Mixing both laws with the same reset distribution gives

$$\mathrm{KL}(sK_0 + \lambda\nu \,\|\, sK_1 + \lambda\nu) \leq s\,\mathrm{KL}(K_0\|K_1) \leq 64st^2 \leq 64t^2. \tag{269}$$

The first inequality follows by retaining the mixture indicator and then using data processing. It is only a divergence bound; the execution rule is not supplied with that indicator. Nonprobe rows contribute zero.

Let $\mathbb{P}^{(k)}_\sigma$ be the law of the decision history through (S_k, A_k), including the algorithm's private random stream U. The law of U and the initial state agree between models. Conditional on a common history, the algorithm uses the same action rule. Thus the adaptive chain rule sums only transition divergences; for $k < T_0$,

$$\mathrm{KL}(\mathbb{P}^{(k)}_0 \| \mathbb{P}^{(k)}_1) \leq 64T_0t^2 \leq \frac{1}{32}, \qquad \mathrm{TV}(\mathbb{P}^{(k)}_0, \mathbb{P}^{(k)}_1) \leq \frac{1}{8}. \tag{270}$$

No stationary-policy restriction has entered this comparison.

Every transition enters j with probability at most

$$\frac{\lambda}{n} + s\theta_+ \leq \frac{\lambda}{4} + (1-\lambda)\frac{5}{32} \leq \frac{13}{64} < \frac{1}{4}.$$

Since the process starts outside j, a non-j decision has probability at least $3/4$ at every time. Let $B^{(k)}_\sigma$ be the indicator that $S_k \neq j$ and A_k is not model σ's larger probe. As functions of a common history and action, $B^{(k)}_0 + B^{(k)}_1 \geq \mathbf{1}\{S_k \neq j\}$. Using (270) to compare expectations gives

$$\frac{\mathbb{E}_0 B^{(k)}_0 + \mathbb{E}_1 B^{(k)}_1}{2} \geq \frac{1}{2}\left(\frac{3}{4} - \frac{1}{8}\right) \geq \frac{1}{4}, \qquad k < T_0. \tag{271}$$

For an arbitrary history-dependent execution Π, write J_σ^Π for its return conditional on U. Arrival telescoping gives

$$a\mathbb{E}_U(J_\sigma^* - J_\sigma^\Pi)(x_1) = a\mathbb{E}_\sigma \sum_{k\geq 0} \gamma^k G_\sigma(S_k, A_k). \tag{272}$$

This follows by using $W_\sigma^* = r + \gamma J_\sigma^*$ in the conditional expectation of each summand; the terminal remainder vanishes because values are bounded and $\gamma < 1$. The identity therefore permits updates based on observed states, rewards, and failed probes. All summands are nonnegative. Since $a\sum_{k<T_0}\gamma^k = 1 - \gamma^{T_0} \geq 1 - e^{-1} \geq 1/2$, equations (267) and (271) imply

$$\frac{1}{2}\sum_{\sigma=0}^{1} a\mathbb{E}_U(J_\sigma^* - J_\sigma^\Pi)(x_1) \geq \frac{2st}{D}\cdot\frac{1}{4}\cdot\frac{1}{2} \geq \frac{st}{5}. \tag{273}$$

Here $\mathbb{E}_\sigma$ averages private randomness and environment transitions, whereas environment transitions are already averaged in J_σ^Π. Thus one fixed model has at least this expected loss. The environment is not selected after observing U or the trajectory.

L.6 Noise calibration and completion of the theorem

For arbitrary $\delta > 0$, use the family above with

$$\begin{aligned} t_{\mathrm{del}} &= \min\left\{\frac{3\delta H}{sd_\lambda}, \frac{1}{64\sqrt{H}}\right\}, \\ t_{\mathrm{full}} &= \min\left\{\frac{\delta H}{s}, \frac{1}{64\sqrt{H}}\right\}. \end{aligned} \tag{274}$$

Both choices satisfy (268) and the probability range in (259), since $H \geq 2$. By (264), the midpoint errors are $asd_\lambda t_{\mathrm{del}}/3$ for deletion and ast_{full} for the full table. Both are at most δ, including when the testing cap is active. Equation (273) now gives

$$\begin{aligned} \mathcal{D}^{\mathrm{exec}}_{\mathrm{full},\lambda}(\delta;\gamma) &\geq \min\left\{\frac{\delta H}{5}, \frac{1-\lambda}{320\sqrt{H}}\right\}, \\ \mathcal{D}^{\mathrm{exec}}_{\mathrm{del},\lambda}(\delta;\gamma) &\geq \min\left\{\frac{3\delta H}{5d_\lambda}, \frac{1-\lambda}{320\sqrt{H}}\right\}. \end{aligned} \tag{275}$$

The two calibrations use two pairs in the same class, addressing two different risks at the same noise level. At $\delta = 0$ both lower bounds are zero, and the upper bounds also vanish. Relabeling j covers any prespecified omission, and the construction already includes every $n \geq 4$ and $A \geq 5$.

In the lower experiment, the reward, root, candidate models, and common input table can all be specified before execution. Even revealing the candidate pair and its common reset distribution leaves the model bit unknown. The midpoint and the two environments are fixed before private randomness. Extra states and actions provide no observation outside the complete-table and history comparisons above. Combining (275) with (245), (246), and (256) gives the precise upper bounds

$$\begin{aligned} \mathcal{D}^{\mathrm{exec}}_{\mathrm{full},\lambda}(\delta;\gamma) &\leq \min\{16n\delta H,\ B_{n,nA}(H),\ 1-\lambda\}, \\ \mathcal{D}^{\mathrm{exec}}_{\mathrm{del},\lambda}(\delta;\gamma) &\leq \min\{8(n-1)\delta H/d_\lambda,\ B_{n,nA}(H),\ 1-\lambda\}. \end{aligned} \tag{276}$$

For each known error budget δ and promised class parameter λ, choose the smaller certificate from the stationary critic decoder and the online execution rule. This is a pointwise minimax choice; it does not prove an unknown-noise adaptive hybrid. The cap $1-\lambda$ applies to either rule. Since $1-\lambda \geq 1/2$, (275) is at least $1/640$ times $\min\{\delta H, H^{-1/2}\}$ for the full library and $1/640$ times $\min\{\delta H/d_\lambda, H^{-1/2}\}$ for deletion. The uniform envelope (257) bounds each upper bound by a constant times $\sqrt{\log(eH)}$ times its corresponding minimum. This proves Theorem 7, with constants independent of δ, H, λ for fixed n, A.

Both testing lower bounds are linear in δ in the following small-noise regime:

$$0 \leq \delta \leq \frac{d_\lambda}{384H^{3/2}}. \tag{277}$$

Indeed, before clipping, $3\delta H/(sd_\lambda) \le 1/(128s\sqrt{H}) \le 1/(64\sqrt{H})$, and the full separation is no larger. The lower bounds in this regime are $\delta H/5$ and $3\delta H/(5d_\lambda)$. At larger noise, the capped experiment and the online upper describe what can be learned during execution.

The coefficient scale is also explicit in this family. From (262), $C_\beta = I - \beta e_j w$, so $(c_\beta)_j = d_\lambda$ and $(c_\beta)_i = 1$ for $i \neq j$. Equations (242) and (265) give $(c_P)_j = a$, $(C_P W_P^*)_j = 1$, and $(C_P W_P^*)_i = x_\beta$ for every $i \neq j$. Thus the deleted expansion has $k = H$ and

$$\alpha_i = x_\beta - \frac{1}{d_\lambda} = -\frac{1}{d_\lambda(1+\beta\theta_+)}, \qquad \|\alpha\|_1 = \frac{n-1}{d_\lambda(1+\beta\theta_+)}. \tag{278}$$

This is a legal bounded-span policy value, including when all transition entries are positive. The proved lower bounds concern execution; one erroneous first probe alone has an additional factor a after normalization. The theorem makes no matching first-action claim, and its bounds do not match dimension, action, or logarithmic factors.

M Fixed-library recovery and common-policy completion

For a fixed common action count A, let P range over all stochastic kernels on the known state space, without a diameter restriction. Define the fixed-library recovery risk

$$\mathcal{R}^{\mathrm{all}}_{\mathcal{C},n,A}(\delta;\gamma) = \inf_D \sup_P \sup_{\|\widetilde{\bar{q}} - \bar{q}_P\|_{\max} \le \delta} \max_{s,u} \|D(\widetilde{\bar{q}})(\cdot \mid s,u) - P(\cdot \mid s,u)\|_1. \tag{279}$$

The infimum is over deterministic decoders D that output probability rows. We abbreviate this risk as $\mathcal{R}^{\mathrm{all}}_{\mathcal{C}}(\delta)$ when n, A, γ are fixed. Upper bounds also hold for arbitrary finite A_s; lower bounds specify the actions they need. Unless dependence is explicitly indicated, Θ constants are numerical and uniform in the displayed parameters. These are precision bounds, not sample-complexity bounds, and the infimum is over decoders for the specified library, not over libraries.

To decode arbitrary noisy tables, first set $\widetilde{q} = H\widetilde{\bar{q}}$ and $J = S \setminus \{j\}$ for the omitted label j. Construct $\widetilde{M}_J$ by the Bellman minimum and fit each probability row by minimizing $\|p\widetilde{M}_J - \widetilde{q}_J(s,u)\|_\infty$ over the simplex. The resulting linear program needs no noise threshold and no inverse of a noisy matrix. Bellman inversion already uses simplex fitting (Letcher et al., 2026); the task-specific guarantee here gives error at most $\min\{2, 4(n-1)(1+\gamma)\delta H^2\}$. Using all columns and $\|C\|_\infty \le 1+\gamma$ instead gives $\min\{2, 2n(1+\gamma)^2\delta H\}$.

For $n \ge 2$ and one common continuation policy π, normalized critics satisfy $\sum_g \bar{q}^\pi_g(s,u) = 1$. Filling a missing column by $1 - \sum_{g \neq j} \widetilde{\bar{q}}^\pi_g$ raises the error from δ to at most $(n-1)\delta$, independently of H. This deterministic completion can precede *any* downstream decision rule, including learning during execution; it does not require knowing π. A full-input rule can also ignore a column. With the same policy convention, model class and loss, the corresponding minimax risks therefore obey

$$\mathfrak{R}^\pi_{\mathrm{full}}(\delta) \le \mathfrak{R}^\pi_{\mathrm{del}}(\delta) \le \mathfrak{R}^\pi_{\mathrm{full}}((n-1)\delta). \tag{280}$$

These are common-policy inputs, not the separately optimized critics above. The conservation identity is classical; the comparison isolates what separate optimization changes. It removes a direct way to recover the redundant column, allowing the additional horizon loss in (292).

N Policy-value coordinates and planning from noisy critics

This section supplies the policy-value coordinates and fitting bounds used by the fixed-omission control proof. Throughout, $n \ge 2$, $0 \le \gamma < 1$, $a = 1-\gamma$, and $H = a^{-1}$. Upper bounds allow arbitrary finite nonempty action sets. The observation consists of all retained normalized minimum-visit critic columns, with arbitrary entrywise error at most δ. For a downstream reward $r : S \to [0,1]$ and a stationary policy π, write

$$W^\pi_P = (I - \gamma P_\pi)^{-1} r, \qquad J^\pi_P = P_\pi W^\pi_P, \qquad W^\pi_P = r + \gamma J^\pi_P. \tag{281}$$

Thus W^π_P counts the current-state reward, whereas J counts rewards starting at the next state. We normalize arrival returns by a. An executed policy may use the history of its evaluated trajectory; it receives no preliminary interaction with the environment.

For an execution policy Π, define its normalized loss by

$$\mathcal{L}_P^{\mathrm{exec}}(\Pi; r, o) = a[J_P^*(o) - J_P^{\Pi}(o)],$$

where J_P^{Π} is the expected discounted arrival return, including when Π is history dependent. The corresponding fixed-library risk is

$$\mathcal{D}_{\mathcal{C}}^{\mathrm{exec}}(\delta; \gamma) = \inf_{\mathcal{A}} \sup_{P} \sup_{\|Z - \bar{q}_{P,\mathcal{C}}\|_{\max} \le \delta} \sup_{r \in [0,1]^n,\, o \in S} \mathbb{E}_U \mathcal{L}_P^{\mathrm{exec}}(\mathcal{A}(Z, r, o; U); r, o). \tag{282}$$

The expectation is over private randomization; the instance (P, Z, r, o) is fixed outside it. An algorithm given the full table can discard a column, giving $\mathcal{D}_{\mathrm{full}}^{\mathrm{exec}} \le \mathcal{D}_{\mathrm{del}}^{\mathrm{exec}}$.

Value-equivalence results already transfer controlled value errors to policy-performance bounds (Grimm et al., 2021; 2022). The task here is to verify a uniform coefficient bound for this fixed avoidance library, including after any one column is deleted. We give the arrival-value comparison explicitly to include $\gamma = 0$.

N.1 Bounded coordinates for every policy value

Let M collect all singleton minimum-occupation columns, so that

$$M_{sg} = \mathbf{1}\{s = g\} + \gamma \min_u q_g(s, u), \qquad q_g(s, u) = P(\cdot \mid s, u) M_g. \tag{283}$$

In particular, $M \ge I$ entrywise. The following argument uses a classical inverse-positive matrix factorization; it does not introduce a new matrix-positivity criterion (Moré, 1970).

Lemma 5 (Policy-value coordinates)**.** *Set $C = M^{-1}$ and $c = C\mathbf{1}$. For every stationary, possibly randomized policy π, with $N_\pi = (I - \gamma P_\pi)^{-1}$,*

$$CN_\pi \ge 0, \qquad H^{-1}\mathbf{1} \le c \le \mathbf{1}, \qquad 0 \le CW_P^\pi \le Hc \quad (0 \le r \le \mathbf{1}). \tag{284}$$

For any omitted coordinate j, put $J = S \setminus \{j\}$. Every such policy value admits

$$W_P^\pi = k\mathbf{1} + M_J \alpha, \qquad 0 \le k \le H, \qquad \|\alpha\|_1 \le H \sum_{i \in J} c_i \le (n-1)H. \tag{285}$$

For the full library it admits the same representation with $k = H/2$ and coefficient norm at most $nH/2$.

Proof. Bellman minimality holds against every stationary policy: $M \le I + \gamma P_\pi M$. Hence

$$B = (I - \gamma P_\pi)M = I - R, \qquad R \ge 0.$$

The positive row $u^\top = \mathbf{1}^\top N_\pi$ satisfies $u^\top B = \mathbf{1}^\top M > 0$, so $u^\top R < u^\top$. Consequently

$$\rho(R) \le \max_i \frac{(u^\top R)_i}{u_i} < 1.$$

Indeed, the right-hand side equals the maximum column sum of

$$\mathrm{diag}(u) R \, \mathrm{diag}(u)^{-1} \ge 0.$$

The Neumann series gives $B^{-1} = \sum_{t \ge 0} R^t \ge I$. This also proves M invertible, and

$$CN_\pi = B^{-1} \ge 0, \qquad c = H^{-1} B^{-1} \mathbf{1} \ge H^{-1}\mathbf{1}.$$

Since $Mc = \mathbf{1}$, $M \ge I$, and $c > 0$, also $c \le \mathbf{1}$. The nonnegative matrix CN_π has row sums Hc; applying it to $0 \le r \le \mathbf{1}$ proves the last inequality in (284). This argument does not require the stronger off-diagonal sign structure of C.

Write $\beta = CW_P^\pi$, set $k = \beta_j / c_j$, and set $\alpha_i = \beta_i - k c_i$ for $i \ne j$. Because $Mc = \mathbf{1}$, this gives (285). The bounds follow from $0 \le \beta_i \le Hc_i$: both k and β_i / c_i lie in $[0, H]$, so $|\alpha_i| \le Hc_i$. For the full library, use

$$W_P^\pi = \frac{H}{2}\mathbf{1} + M\left(\beta - \frac{H}{2}c\right), \qquad \left\|\beta - \frac{H}{2}c\right\|_1 \le \frac{H}{2}\sum_i c_i \le \frac{nH}{2}.$$

□

The lemma holds simultaneously for all stationary policies of the true model. In particular, it covers a policy selected using a different fitted model and then evaluated in the true model. It does not assert that a common policy generates the separate columns of M.

N.2 A CONSTRUCTIVE PLANNING UPPER BOUND

Let $\mathcal{I}$ be either S or $S \setminus \{j\}$, and let Z be the observed normalized table on these columns. Set

$$\widetilde{q} = HZ, \qquad \widetilde{M}_{sg} = \mathbf{1}\{s = g\} + \gamma \min_u \widetilde{q}_g(s,u), \quad g \in \mathcal{I}.$$

For every state-action row, choose

$$\widehat{p}_{su} \in \arg\min_{z \in \Delta_n} \|z\widetilde{M} - \widetilde{q}(s,u,\cdot)\|_\infty, \qquad \Delta_n = \{z \geq 0 : z\mathbf{1} = 1\}. \tag{286}$$

This is the ℓ_∞ version of the simplex fit in Lemma 6; an ℓ_1 fitting objective is not being assumed here. It is a finite linear program, with a minimizer for every real input table. Fix a deterministic tie rule. The fitted rows form a stochastic model $\widehat{P}$, without requiring a noisy inverse, a continuation-policy label, the omitted column, or knowledge of δ.

The minimum is 1-Lipschitz in the supremum norm, giving $\|\widetilde{M} - M_{\mathcal{I}}\|_{\max} \leq \gamma H\delta$. The true row $p = p_{su}$ has fitting residual at most $(1+\gamma)H\delta$, so the minimizing row has no larger residual. It follows directly that

$$\begin{aligned}\|(\widehat{p} - p)M_{\mathcal{I}}\|_\infty &\leq \|\widehat{p}(M_{\mathcal{I}} - \widetilde{M})\|_\infty + \|\widehat{p}\widetilde{M} - \widetilde{q}(s,u,\cdot)\|_\infty \\ &\quad + \|\widetilde{q}(s,u,\cdot) - q(s,u,\cdot)\|_\infty \\ &\leq 2(1+\gamma)H\delta =: \varepsilon.\end{aligned} \tag{287}$$

Both rows sum to one. Constants therefore disappear when their difference acts on the representation in Lemma 5. Uniformly over all rows, stationary policies, and $r : S \to [0,1]$,

$$|(\widehat{p}_{su} - p_{su})W_P^\pi| \leq e_{\mathcal{I}}, \qquad e_{\mathcal{I}} = \begin{cases} n(1+\gamma)\delta H^2, & \mathcal{I} = S, \\ 2(n-1)(1+\gamma)\delta H^2, & \mathcal{I} = S \setminus \{j\}. \end{cases} \tag{288}$$

The unknown coordinates in the lemma are only used for this analysis. The fit itself is independent of r, so one fitted model supports every downstream reward, including rewards revealed after fitting.

For any stationary policy, (281) gives the exact arrival-value comparison

$$J_P^\pi - J_{\widehat{P}}^\pi = (I - \gamma\widehat{P}_\pi)^{-1}(P_\pi - \widehat{P}_\pi)W_P^\pi. \tag{289}$$

To see this, subtract $P_\pi W_P^\pi$ and $\widehat{P}_\pi W_{\widehat{P}}^\pi$, then use $W_P^\pi - W_{\widehat{P}}^\pi = \gamma(J_P^\pi - J_{\widehat{P}}^\pi)$. The tested vector on the right is a true-model policy value, so (288) applies even when π was chosen in $\widehat{P}$. The fitted resolvent has norm H; averaging action rows according to π preserves the residual bound. Thus

$$a\|J_P^\pi - J_{\widehat{P}}^\pi\|_\infty \leq e_{\mathcal{I}}. \tag{290}$$

Choose $\widehat{\pi}$ greedily with respect to the fitted optimal arrival action values $\widehat{P}_{su}W_{\widehat{P}}^*$, and let π^* be an optimal policy in P. Optimality in the fitted model and (290) give, componentwise,

$$\begin{aligned}a(J_P^* - J_P^{\widehat{\pi}}) &= a(J_P^{\pi^*} - J_{\widehat{P}}^{\pi^*}) + a(J_{\widehat{P}}^{\pi^*} - J_{\widehat{P}}^{\widehat{\pi}}) + a(J_{\widehat{P}}^{\widehat{\pi}} - J_P^{\widehat{\pi}}) \\ &\leq 2e_{\mathcal{I}}\mathbf{1}.\end{aligned}$$

Normalized returns lie in $[0,1]$, so the execution loss is at most $\min\{1, 2e_{\mathcal{I}}\}$. At $\gamma = 0$, (289) reduces to $J_P^\pi - J_{\widehat{P}}^\pi = (P_\pi - \widehat{P}_\pi)r$. The greedy convention selects actions maximizing $\widehat{p}_{su}r$. An arbitrary policy tied for current-state optimality would not suffice at this endpoint, since every current-state value then equals r.

O EXACT REDUNDANCY AND NOISY RECOVERY

Let $\mathcal{R}_{\mathcal{C}}^{\mathrm{all}}(\delta)$ be the minimax maximum-row ℓ_1 transition error over unrestricted finite kernels, with decoders constrained to output probability rows and the same noisy interface; Appendix M gives the formal definition. The inverse structure determines both task count and the precision cost of deletion. Let $\mathcal{C}_{\mathrm{full}} = \{e_g : g \in S\}$ and let $\mathcal{C}_{\mathrm{del}}$ omit any one prespecified singleton cost.

Corollary 6 (Complete-library decoding)**.** *Complete minimum-visit critics determine every finite discounted kernel:*

$$P_{sa} = q(s, a, \cdot)M^{-1}. \tag{291}$$

Theorem 8 (Task count and precision)**.** *For stochastic transition rows, any $n-1$ singleton-avoidance tasks identify every kernel. Fewer than $n-1$ arbitrary fixed state rewards cannot do so. For $n \geq 3$ and a common action count $A \geq 2$,*

$$\begin{aligned} \mathcal{R}^{\mathrm{all}}_{\mathrm{full}}(\delta) &= \Theta(\min\{1, n\delta H\}), \\ \mathcal{R}^{\mathrm{all}}_{\mathrm{del}}(\delta) &= \Theta(\min\{1, n\delta H^2\}). \end{aligned} \tag{292}$$

Both upper bounds hold for arbitrary finite action sets. With two states, the one-task rate is instead $\Theta(\min\{1, \delta H\})$ on the class $A \geq 1$. Under one common known continuation policy, deleting a singleton admits $O(\min\{1, n\delta H\})$ error for every n.

For exact identification, write $C = M^{-1}$, omit coordinate j, and put $\mathcal{I} = S \setminus \{j\}$. A difference η between probability rows obeys $\eta\mathbf{1} = 0$. If $\eta M_{\mathcal{I}} = 0$, then $\eta M = te_j^\top$, so $0 = \eta\mathbf{1} = t(C\mathbf{1})_j$. Theorem 2 gives $(C\mathbf{1})_j \geq H^{-1} > 0$, forcing $\eta = 0$. For the converse, fewer than $n-1$ reward vectors leave a nonzero direction orthogonal to them and to $\mathbf{1}$. Two distributions along that direction, used as every transition row, produce identical complete critics.

With error, however, reconstructing the missing coordinate may divide by $(C\mathbf{1})_j$, as small as H^{-1}. Appendix P proves the sharp restricted inverse bound

$$\|\eta\|_1 \leq 2H\|\eta M_{\mathcal{I}}\|_1, \qquad \eta\mathbf{1} = 0. \tag{293}$$

A simplex fit to the noisy Bellman equations gives these uniform rates without inverting a noisy matrix. Its row objective can be either a maximum or a sum of absolute residuals; the proofs specify which norm they use. Appendix M records the decoder and constants.

The additional horizon factor arises when the omitted singleton reports an immediate visit, while retained singletons see only a one-step delay before the same distribution is reached. Their normalized signal is smaller by H, although the transition change is the same. Appendix P.4 gives the complete-critic construction and its matching lower-bound proof.

Under a common continuation policy, normalized singleton critics sum to one. Filling a missing column therefore multiplies its error by at most $n-1$, independently of H. This completion can precede any downstream rule, including online learning. Separate optimization breaks that identity; Appendix M states the formal information comparison.

P TASK BUDGETS AND THE PRICE OF DELETING ONE SINGLETON

This section proves Theorem 8, including the two-state exception. Throughout, $H = (1-\gamma)^{-1}$ and $0 \leq \gamma < 1$. Costs are minimized. For a cost c, the current-state value and exported arrival-cost critic obey

$$V_c = c + \gamma \min_u q_c(\cdot, u), \qquad q_c(s, u) = P(\cdot \mid s, u)V_c, \qquad \bar{q}_c = q_c/H.$$

The observation consists of the complete table, with entrywise error at most δ. We use the deterministic adversarial risk in (279), abbreviating it as $\mathcal{R}^{\mathrm{all}}_{\mathcal{C}}(\delta)$ when n, A, γ are fixed. The decoder knows the costs and interface, but no transition rows or continuation policies. Its loss is the maximum transition-row ℓ_1 error.

All upper bounds below allow arbitrary finite nonempty action sets. For $n \geq 3$, the deleted-library lower bound uses two actions per state, so its matching rate holds for every fixed common action count $A \geq 2$. The full-library rate and the two-state lower bound already hold with one action. Duplicating an action changes none of the values or arguments. The stated extra horizon factor is not asserted for a prescribed single-action class.

P.1 THE EXACT UNIVERSAL TASK COUNT

Proposition 7 (Minimum number of fixed state-cost tasks)**.** *For $n \geq 2$, any fixed family of fewer than $n-1$ scalar state costs fails to identify every stochastic transition kernel from its complete optimal critic tables. Conversely, any $n-1$ of the n singleton minimum-visit tasks identify every such kernel.*

Proof. Fix costs $c_1, \dots, c_m \in \mathbb{R}^n$ with $m < n-1$. There is a nonzero row η such that

$$\eta \mathbf{1} = 0, \qquad \eta c_g = 0 \quad (1 \le g \le m).$$

Write $t = \|\eta_+\|_1 = \|\eta_-\|_1 > 0$ and $\nu_\pm = \eta_\pm / t$. In model $P_\pm$, set every state-action row equal to $\nu_\pm$. Actions have no effect on the future law, and

$$q_{c_g}^{P_\pm}(s,u) = H \nu_\pm c_g.$$

Thus the complete tables agree for every cost, although every transition row has ℓ_1 distance two. In fact, every decoder has worst-case error at least one even at $\delta = 0$. This argument allows arbitrary fixed real state costs, not only singleton costs. It also works with any prescribed action count by duplicating rows.

For sufficiency, let M be the full minimum-occupation matrix and put $C = M^{-1}$. Theorem 2 gives

$$C_{ii} > 0, \quad C_{ij} \le 0 \ (i \ne j), \quad C_{ii} \le 1, \qquad c := C\mathbf{1} \ge H^{-1}\mathbf{1}. \tag{294}$$

In particular, $H^{-1} \le c_i \le 1$. Fix the omitted state j and write $J = S \setminus \{j\}$. The retained critics determine

$$(M_J)_{sg} = \mathbf{1}\{s = g\} + \gamma \min_u q_g(s,u), \quad g \in J.$$

Each transition row p solves $pM_J = q_J(s,u)$ and $p\mathbf{1} = 1$. To show uniqueness, suppose $\eta M_J = 0$ and $\eta \mathbf{1} = 0$. Then $\eta M = t e_j^\top$ for some $t \in \mathbb{R}$, so $\eta = t C_j$, where C_j is row j of C. Therefore $0 = \eta \mathbf{1} = t c_j$, and $c_j \ge H^{-1} > 0$ implies $\eta = 0$. Equivalently, $[M_J, \mathbf{1}]$ is nonsingular. □

For $n = 1$ the transition row is already known, so zero tasks suffice. The stochastic-row constraint in Proposition 7 is essential: the missing equation is supplied by $p\mathbf{1} = 1$.

P.2 A SIMPLEX DECODER FOR ARBITRARY NOISY TABLES

We first record a residual argument used for both singleton and overlapping libraries. It does not require the noisy table to be the optimal critics of any MDP.

Lemma 6 (Residual decoding)**.** *Let $V = [V_{c_1}, \dots, V_{c_m}]$ be a true current-state value matrix. Suppose, for $r \in \{1, \infty\}$, every zero-sum row satisfies*

$$\|\eta\|_1 \le K \|\eta V\|_r. \tag{295}$$

Given an arbitrary normalized table $\widehat{\widehat{q}}$, set $\widehat{q} = H\widehat{\widehat{q}}$ and

$$\widehat{V}_{sg} = c_g(s) + \gamma \min_u \widehat{q}_g(s,u).$$

For each state-action pair, choose

$$\widehat{p} \in \arg\min_{p \in \Delta_n} \|p\widehat{V} - \widehat{q}(s,u,\cdot)\|_r, \qquad \Delta_n = \{p \in \mathbb{R}^n : p \ge 0,\ p\mathbf{1} = 1\}. \tag{296}$$

If

$$\varepsilon_r = \max_s \left\| \left(\max_u |\widehat{q}_g(s,u) - q_g(s,u)| \right)_{g=1}^m \right\|_r,$$

then the resulting stochastic kernel satisfies

$$\max_{s,u} \|\widehat{p}_{su} - p_{su}\|_1 \le \min\{2,\ 2K(1+\gamma)\varepsilon_r\}. \tag{297}$$

Entrywise normalized error δ gives $\varepsilon_1 \le mH\delta$ and $\varepsilon_\infty \le H\delta$.

Proof. The scalar minimum is 1-Lipschitz in the supremum norm. Consequently every row of $E = \widehat{V} - V$ has ℓ_r norm at most $\gamma\varepsilon_r$. Because a probability row forms a convex combination of these rows, $\|pE\|_r \le \gamma\varepsilon_r$ for every $p \in \Delta_n$. The true row has residual at most $(1+\gamma)\varepsilon_r$ in (296); the minimizing row has no larger residual. Using $pV = q(s,u,\cdot)$,

$$\begin{aligned} \|(\widehat{p} - p)V\|_r &\le \|\widehat{p}(V - \widehat{V})\|_r + \|\widehat{p}\widehat{V} - \widehat{q}(s,u,\cdot)\|_r + \|\widehat{q}(s,u,\cdot) - q(s,u,\cdot)\|_r \\ &\le 2(1+\gamma)\varepsilon_r. \end{aligned}$$

The row $\widehat{p} - p$ sums to zero, so (295) applies. The probability constraints also bound its norm by two.

The objective in (296) is continuous on the compact simplex, so a minimizer exists for every real input table. For $r = 1$, introduce one nonnegative absolute-residual slack per task; for $r = \infty$, introduce a single common residual bound. Both programs are linear. Neither invertibility of $\widehat{V}$ nor knowledge of δ is required. □

The same calculation compares two legal models without fitting a row. If their raw critic difference has mixed norm ε_r as above, then

$$\max_{s,u} \|p^0_{su} - p^1_{su}\|_1 \le \min\{2, K(1+\gamma)\varepsilon_r\}. \tag{298}$$

Indeed, $(p^0 - p^1)V^0 = (q^0 - q^1)(s, u, \cdot) + p^1(V^1 - V^0)$.

P.3 The full singleton library

The bound $\|M^{-1}\|_\infty \le 1 + \gamma$ from Theorem 2 implies $\|\eta\|_1 \le (1+\gamma)\|\eta M\|_1$. Lemma 6 therefore gives

$$\mathcal{R}^{\mathrm{all}}_{\mathcal{C}_{\mathrm{full}}}(\delta) \le \min\{2,\, 2n(1+\gamma)^2 H\delta\}. \tag{299}$$

For $n \ge 2$ and $A \ge 3$, consider the following complete-table lower bound. Give every state a stay action and an action leading to its successor on a fixed cycle. Add a third, probe action. Only the probe row at a distinguished state o will differ between models; the other probe rows may duplicate their stay actions. For target g, a state other than g can avoid it forever by staying. From g, the controller can leave on the cycle and then stay elsewhere. Hence the minimum continuation occupation is zero at every state and $M = I$, independently of the distinguished probe row. In particular, the full exported raw singleton table is exactly $q = P$.

Put $N = 2\lfloor n/2 \rfloor$ and choose N states split into two equal sets. Assign signs $\sigma_i \in \{-1, 1\}$ with equal numbers of each sign. For $0 \le t \le 1$, use the two probe distributions

$$p_\pm(i) = \frac{1 \pm t\sigma_i}{N} \quad \text{on these } N \text{ states}, \qquad p_\pm(i) = 0 \quad \text{elsewhere}.$$

Their ℓ_1 distance is $2t$. Every critic entry outside this probe agrees, and the complete normalized table distance is $2t/(NH)$. Choose $t = \min\{1, NH\delta\}$. The entrywise midpoint of the two complete tables lies within δ of each. A decoder receives the same permissible input in the two models, and the triangle inequality forces error at least t on one probe row. Thus

$$\mathcal{R}^{\mathrm{all}}_{\mathcal{C}_{\mathrm{full}}}(\delta) \ge \min\{1,\, 2\lfloor n/2 \rfloor H\delta\}. \tag{300}$$

Together with (299), this gives $\Theta(\min\{1, nH\delta\})$ with absolute constants, uniformly over $n \ge 2$, $\gamma \in [0, 1)$, on classes containing this three-action family.

For completeness, when $n \ge 3$ the same order also has a one-action construction. Reserve a root o and take $N' = 2\lfloor (n-1)/2 \rfloor$ other states as the support of the two balanced probe laws above. Choose an anchor in this support, send every non-root state to that anchor, and vary only the root row. No row returns to the root. For a singleton cost, the two root arrival critics differ by the corresponding coordinate of $p_+ - p_-$: after the first arrival their continuations agree. All other critics agree. Replacing N by N' therefore proves the lower bound for every fixed $A \ge 1$ when $n \ge 3$. We do not use this observation to assert an extra horizon factor for the deleted library with one action.

P.4 A sharp restricted inverse after deleting one singleton

Lemma 7. *Let $J = S \setminus \{j\}$. Every zero-sum row satisfies*

$$\|\eta\|_1 \le 2H\|\eta M_J\|_1. \tag{301}$$

For $n \ge 3$, the coefficient $2H$ is attained by a legal MDP for each $\gamma \in [0, 1)$.

Proof. Use C, c from (294). Write $z = \eta M$ and $u = z_J$. Since $0 = \eta \mathbf{1} = zc$,

$$z_j = -\sum_{i \in J} u_i \frac{c_i}{c_j}, \qquad \eta = \sum_{i \in J} u_i \left(C_i - \frac{c_i}{c_j} C_j \right).$$

If $d_i = C_{ii}$, the sign pattern gives $\|C_i\|_1 = 2d_i - c_i$. Hence

$$\begin{aligned} \left\| C_i - \frac{c_i}{c_j} C_j \right\|_1 &\le (2d_i - c_i) + \frac{c_i}{c_j}(2d_j - c_j) \\ &= 2d_i + 2\frac{c_i}{c_j} d_j - 2c_i \\ &\le 2 + 2c_i(c_j^{-1} - 1) \le 2H. \end{aligned}$$

The last line uses $d_i, d_j \le 1$, $0 < c_i \le 1$, and $H^{-1} \le c_j \le 1$. Summing with weights $|u_i|$ proves the bound. The construction below establishes sharpness. □

Taking $m = n - 1$, $r = 1$, and $K = 2H$ in Lemma 6, or $r = \infty$ and $K = 2(n-1)H$ using $\|x\|_1 \le (n-1)\|x\|_\infty$, yields

$$\mathcal{R}^{\mathrm{all}}_{\mathcal{C}_{\mathrm{del}}}(\delta) \le \min\{2,\, 4(n-1)(1+\gamma)H^2\delta\}. \tag{302}$$

To prove a matching complete-table lower bound, take $n = N + 2$, $N \ge 1$, with states $\{j, r_1, \dots, r_N, z\}$. The omitted target is j, and w is uniform on the r_i. Every action at j has row w, and z is absorbing. Each r_i with $i \ge 2$ has a stay action and an action to z. State r_1 has exactly two actions: an action to z and a probe with transition row

$$p_\theta = \theta e_j^\top + (1-\theta) w, \qquad 0 \le \theta \le 1.$$

Duplicate the rows at j and z to obtain a common count of two. For every θ, the entire current-state minimum-occupation matrix has rows

$$M_j = e_j^\top + \gamma w, \qquad M_{r_i} = e_{r_i}^\top, \qquad M_z = H e_z^\top. \tag{303}$$

To verify these rows, first consider a target other than z. At r_1, the action to z has zero continuation value. At any other r_i, staying has zero continuation value unless the target is r_i itself, in which case the action to z has zero continuation value. For target z, staying suffices at every r_i with $i \ge 2$. At r_1, the probe has zero continuation value under the proposed column: it stays within $\{j, r_1, \dots, r_N\}$, where that column is zero. All competing continuations are nonnegative. Thus the minimum continuation is zero at every r_i for every target. The forced row w at j then gives $M_j = e_j^\top + \gamma w$, and absorption gives $M_z = H e_z^\top$. These are all Bellman equations, so uniqueness proves (303) for every θ. When $N = 1$, the same verification applies: repeated probes and the forced transition from j remain in $\{j, r_1\}$ and avoid z.

Since $q = PM$, only the probe row of the complete critic table changes. Its raw entries are

$$q_j(r_1, \mathrm{probe}) = \theta, \qquad q_{r_i}(r_1, \mathrm{probe}) = \frac{1 - \theta/H}{N}, \qquad q_z(r_1, \mathrm{probe}) = 0. \tag{304}$$

The omitted task records an immediate visit to j. The retained tasks record a one-step delay before the same distribution w is reached. For parameters separated by t, the transition-row distance is $2t$, while the complete retained normalized table distance is exactly $t/(NH^2)$. Choose $t = \min\{1, 2NH^2\delta\}$ and use the common midpoint table. The same triangle-inequality argument gives

$$\mathcal{R}^{\mathrm{all}}_{\mathcal{C}_{\mathrm{del}}}(\delta) \ge \min\{1,\, 2(n-2)H^2\delta\}. \tag{305}$$

This proves the claimed $\Theta(\min\{1, nH^2\delta\})$ order for $n \ge 3$ and every fixed common action count $A \ge 2$. Relabeling makes the omitted target any preselected state. For the same matrix, $\eta = e_j^\top - w$ has $\|\eta\|_1 = 2$ and $\|\eta M_J\|_1 = H^{-1}$, which proves the sharpness claimed in Lemma 7.

All of these comparisons use the complete critics. In particular, (305) is not inferred from a small singular value of a fixed matrix without constructing the corresponding models. With the omitted target restored, the same pair has normalized table distance t/H: it directly reports whether the probe first reached j.

P.5 THE TWO-STATE EXCEPTION AND THE COMMON-POLICY COMPARISON

For $n = 2$, retain target 1 and write its value column as $(m_1, m_2)^\top$. It has $m_1 > m_2$. Indeed, both entries are nonnegative and the column is nonzero; if state 2 attained its maximum, its zero-cost Bellman equation would give $m_2 \le \gamma \max_s m_s < \max_s m_s$, a contradiction. Set $x_s = \min_u P(1 \mid s, u)$. Since $m_1 > m_2$, minimizing a continuation is the same as minimizing this probability. Subtracting the two Bellman equations gives

$$m_1 - m_2 = \frac{1}{1 - \gamma(x_1 - x_2)} \ge \frac{1}{1+\gamma}.$$

Every zero-sum row is $(t, -t)$, so

$$\|\eta\|_1 \le 2(1+\gamma)|\eta M_{\{1\}}|. \tag{306}$$

The deterministic alternating chain attains this coefficient. Lemma 6 now gives

$$\mathcal{R}^{\mathrm{all}}_{\mathcal{C}_{\mathrm{del}}}(\delta) \le \min\{2, 4(1+\gamma)^2 H\delta\} \qquad (n = 2).$$

For a matching lower bound with one action, let state 2 be absorbing and set $P_\theta(\cdot \mid 1) = \theta e_1^\top + (1-\theta)e_2^\top$ for $0 \le \theta \le 1/2$. For target 1, the current-state value is $(1/(1-\gamma\theta), 0)^\top$, so

$$\bar{q}_1(1) = \frac{H^{-1}\theta}{1-\gamma\theta}, \qquad \bar{q}_1(2) = 0.$$

Compare $\theta = 0$ with $\theta = t = \min\{1/2, H\delta\}$. Since $1 - \gamma t \ge 1/2$, the normalized critic gap is at most $2t/H \le 2\delta$, whereas the transition-row distance is $2t$. The midpoint observation therefore forces worst-case error at least t. Moreover, the one-action identity $\bar{q}_2 = \mathbf{1} - \bar{q}_1$ shows that the same midpoint argument applies to the full singleton table. Consequently, for every fixed $A \ge 1$,

$$\mathcal{R}^{\mathrm{all}}_{\mathcal{C}}(\delta) \ge \min\{1/2, H\delta\} \ge \tfrac{1}{2}\min\{1, H\delta\}, \qquad n = 2, \quad \mathcal{C} \in \{\mathcal{C}_{\mathrm{del}}, \mathcal{C}_{\mathrm{full}}\}.$$

There is no extra horizon factor in this case. The proof covers $\gamma = 0$ without a limiting argument.

It is useful to distinguish separate optimization from a common fixed continuation policy. If all tasks instead follow the same known stationary policy π after the queried first action, then

$$M_\pi = (I - \gamma P_\pi)^{-1}, \qquad C_\pi = M_\pi^{-1}, \qquad C_\pi \mathbf{1} = H^{-1}\mathbf{1}.$$

The omitted-coordinate calculation becomes

$$\eta = \sum_{i \in J} (\eta M_\pi)_i ((C_\pi)_i - (C_\pi)_j), \qquad \|\eta\|_1 \le 2(1+\gamma)\|(\eta M_\pi)_J\|_1.$$

Policy averaging, like taking a minimum, is 1-Lipschitz in each critic column. The same simplex proof therefore gives arbitrary-noise error at most $\min\{2, 4(n-1)(1+\gamma)^2 H\delta\}$ from $n-1$ columns. Equivalently, $\sum_g q^\pi_g = H$ reconstructs the missing column by conservation. This comparison is not an additional policy assumption in the preceding theorems. It explains why their deleted-library H^2 lower bound does not extend to the single-action class.

Q RECOVERY AND CONTROL AFTER COMPLETE ACQUISITION

Compression receives all n noisy singleton critics before choosing a column to discard. Noise is entrywise in the complete normalized arrival-cost table, and the recovery loss is maximum transition-row ℓ_1 error.

Q.1 SELECTING A COLUMN FROM THE SAME NOISY FULL TABLE

Corollary 7 (Noisy selection preserves the full-library horizon order)**.** *For $n \ge 2$, there are an encoder E and decoder D, neither requiring δ, with the following property. Given any full table Z satisfying $\|Z - \bar{q}_P\|_{\max} \le \delta$, the encoder outputs an index j and the unchanged subtable Z_J, where $J = S \setminus \{j\}$. From this pair alone the decoder returns a stochastic kernel satisfying*

$$\max_{s,u} \|\mathsf{D}(\mathsf{E}(Z))(\cdot \mid s, u) - P(\cdot \mid s, u)\|_1 \le \min\{2, 24nH\delta\}.$$

The index is included in the retained column labels, or can be recorded using $\lceil \log_2 n \rceil$ additional bits.

Proof. Write $C = M^{-1}$, $c = C\mathbf{1}$, and $K = 1 + \gamma$. Theorem 2 gives $c_i \geq H^{-1}$ and $\|C\|_\infty \leq K$, where the matrix norm is the induced row-sum norm. First, if $c_j \geq \alpha \max_i c_i$ for some $\alpha > 0$, every zero-sum row η satisfies

$$\eta = \sum_{i \neq j} (\eta M)_i \left(C_i - \frac{c_i}{c_j} C_j \right), \qquad \|\eta\|_1 \leq (1 + \alpha^{-1}) K \|\eta M_J\|_1. \tag{307}$$

Indeed, $(\eta M)c = \eta\mathbf{1} = 0$ determines its missing coordinate, and each row of C has ℓ_1 norm at most K.

The encoder forms

$$\widehat{M}_{sg} = \mathbf{1}\{s = g\} + \gamma H \min_u Z_g(s, u).$$

When $\widehat{M}$ is invertible, it sets $\widehat{c} = \widehat{M}^{-1}\mathbf{1}$ and chooses the smallest index maximizing $\widehat{c}_j$; otherwise it chooses a fixed index. It outputs precisely (j, Z_J). The decoder applies the $r = 1$ simplex program in Lemma 6 to these retained columns. Thus both operations use the same noisy input; selection does not obtain an exact matrix, and the retained table is not replaced by a second observation.

Put $B = \widehat{M} - M$ and $\tau = \gamma K n H \delta$. Taking a minimum is 1-Lipschitz, so $\|B\|_\infty \leq \gamma n H \delta$. If $\tau \leq 1/5$, the Neumann-series bound and $\widehat{c} - c = -\widehat{M}^{-1} B c$ give

$$\|\widehat{M}^{-1}\|_\infty \leq \frac{K}{1 - \tau}, \qquad \|\widehat{c} - c\|_\infty \leq \frac{\tau}{1 - \tau} \max_i c_i.$$

Comparing the maximizers of c and $\widehat{c}$ therefore yields

$$c_j \geq \left(1 - \frac{2\tau}{1 - \tau} \right) \max_i c_i \geq \tfrac{1}{2} \max_i c_i.$$

Equation (307) has coefficient $3K$, and Lemma 6 bounds the error by $6K^2(n-1)H\delta \leq 24nH\delta$. If $\tau > 1/5$, then $\gamma > 0$ and $nH\delta > 1/(5\gamma K) \geq 1/10$. Any stochastic output has error at most $2 \leq 20nH\delta$, including when $\widehat{M}$ is singular. This proves the assertion for arbitrary noise; $\gamma = 0$ belongs to the first case. □

Replacing column j of M by $\mathbf{1}$ multiplies its determinant by c_j. Hence maximizing c_j maximizes the determinant magnitude of the replacement matrix. The assertion above is its noisy, realizable compression guarantee. It retains $n - 1$ columns after accessing all n tasks.

Q.2 Control from the retained columns

Apply the encoder and decoder of Corollary 7 before the downstream reward is disclosed, then compute an optimal arrival-greedy policy in the fitted kernel. Selection and fitting use only the complete noisy table and do not require δ.

Corollary 8 (Control after full acquisition and compression)**.** *For $n \geq 2$, arbitrary finite action sets, $0 \leq \gamma < 1$, and a fixed $0 < \kappa < 1$ with $\kappa(P) \leq \kappa$, this protocol's fitted arrival-greedy optimal policy has normalized execution loss at most $\min\{1, 24nH\delta/(1-\kappa)\}$. Put $m = \sum_s |A_s|$. With the separate online rule of Corollary 5, the loss is at most*

$$\min\{1, 24nH\delta/(1-\kappa), B_{n,m,\kappa}(H)\}.$$

Branch selection may use the declared error budget and class parameters.

Corollary 7 bounds transition error by $\min\{2, 24nH\delta\}$. In the true model, policy values have span at most $(1 - \gamma\kappa)^{-1} \leq (1 - \kappa)^{-1}$. Probability-row differences cancel constants, so (289) gives the control bound without requiring the fitted model to contract. Thus the same retained column count can support different guarantees under different acquisition protocols.

R A common-reset identity

We state the identity for a fixed bounded state reward r, suppressing r in the value-function subscripts. Let K be a stochastic transition kernel, let ν be a probability distribution independent of state and action, and set

$$P = (1 - \lambda)K + \lambda\mathbf{1}\nu, \qquad 0 \leq \lambda \leq 1, \quad 0 \leq \gamma < 1, \quad \beta = \gamma(1 - \lambda). \tag{308}$$

Both processes use continuing arrival rewards.

Lemma 8 (Common reset)**.** *Under* (308)*, with* $V^*_{K,\beta}(s) = \max_a Q^*_{K,\beta}(s,a)$*,*

$$Q^*_{P,\gamma}(s,a) = (1-\lambda)Q^*_{K,\beta}(s,a) + \frac{\lambda}{1-\gamma}\nu\big(r + \beta V^*_{K,\beta}\big)\,. \tag{309}$$

The second term is constant across state-action pairs.

Proof. Write $h = r + \beta V^*_{K,\beta}$ and $k = \gamma\lambda\nu(h)/(1-\gamma)$, where k also denotes the corresponding constant function. The base Bellman equation is $h = r + \beta \max_a K_a h$, and $Q^*_{K,\beta}(s,a) = K_{s,a}h$. Since both kernels preserve constants and $1-\lambda \geq 0$,

$$\begin{aligned} r + \gamma \max_a P_a(h+k) &= r + \beta \max_a K_a h + \gamma\lambda\nu(h) + \gamma k \\ &= h + k. \end{aligned}$$

Uniqueness of the bounded Bellman fixed point therefore identifies $h + k = r + \gamma V^*_{P,\gamma}$. Applying $P_{s,a}$ gives

$$Q^*_{P,\gamma}(s,a) = (1-\lambda)K_{s,a}h + \lambda\nu(h) + k = (1-\lambda)Q^*_{K,\beta}(s,a) + \frac{\lambda\nu(h)}{1-\gamma},$$

which proves the identity. □

For two base kernels with identical complete critics for a reward, their optimal state values and h also coincide. A common reset consequently preserves this exact equality, even if ν reaches states that the base process never revisits. For any reward, all within-state action gaps are multiplied by $1-\lambda$. These statements use the changed discount $\beta = \gamma(1-\lambda)$ and continuing returns; they do not assert the same identity for goal-dependent termination.

S An arbitrary minimum of Green matrices need not suffice

The controlled row-selection structure of Theorem 2 cannot be replaced by an arbitrary collection of fixed-kernel Green matrices. Consider the nonnegative kernels

$$K_1 = \begin{pmatrix} 0 & 1/4 & 1/8 & 0 \\ 0 & 0 & 1/2 & 1/8 \\ 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad K_2 = \begin{pmatrix} 0 & 1/4 & 1/4 & 0 \\ 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 1/4 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

Both have row sums at most $5/8$, so they define ordinary discounted processes with possible killing. They are strictly upper triangular, and their Green matrices $U_b = (I - K_b)^{-1} = I + K_b + K_b^2 + K_b^3$ are

$$U_1 = \begin{pmatrix} 1 & 1/4 & 1/4 & 3/32 \\ 0 & 1 & 1/2 & 1/4 \\ 0 & 0 & 1 & 1/4 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad U_2 = \begin{pmatrix} 1 & 1/4 & 1/4 & 1/8 \\ 0 & 1 & 0 & 1/4 \\ 0 & 0 & 1 & 1/4 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Their entrywise minimum and its inverse are

$$U = \min(U_1, U_2) = \begin{pmatrix} 1 & 1/4 & 1/4 & 3/32 \\ 0 & 1 & 0 & 1/4 \\ 0 & 0 & 1 & 1/4 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \qquad U^{-1} = \begin{pmatrix} 1 & -1/4 & -1/4 & 1/32 \\ 0 & 1 & 0 & -1/4 \\ 0 & 0 & 1 & -1/4 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

The positive off-diagonal entry $(U^{-1})_{14} = 1/32$ means that U is not an inverse M-matrix.

There is no contradiction with the theorem: the two-element collection $\{K_1, K_2\}$ omits independent choices of rows. For example, taking the first row of K_1 and the second row of K_2 gives a kernel whose Green matrix has entry $(1,3)$ equal to $1/8$, below $U_{13} = 1/4$. Allowing the full family of statewise action choices thus changes the lower envelope. Bellman elimination preserves that full row-selection family, which is the additional structure used in the proof.