COMPETENT AGENTS REVEAL EXACTLY WHAT A RANDOM WALK SEES:

THE COMPLETE WORLD-MODEL INFORMATION IN BEHAVIOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Which properties of a partially observed environment must a competent agent's choices reveal? Cifuentes (2026, Section 4.2) poses this identification question for goals that name only observations and actions, and leaves it open. We answer it for finite communicating models that admit a policy optimal for every such goal and every hidden start. The answer is the *random-walk law*, the stationary distribution of observation–action words under independent uniform actions. This law determines every history-dependent controlled prediction, and it is exactly the information guaranteed by competence. An information converse shows that no quantity outside it is determined by every competent policy. A constructive recovery obtains it from sampled actions at one supplied observation, under competence averaged over the queried goals. Both directions extend to a fixed relative loss below one half on the full finite class without a common state cap, where the converse connects equal-law models through a finite common predictive realization. A known state cap changes the answer at positive loss: three deterministic-emission states preserve the characterization, while four can force one additional bit.

1 INTRODUCTION

"the aim of science is not things themselves … but the relations between things"
Henri Poincaré, *Science and Hypothesis* (1905, p. xxiv)

An agent's decisions reveal which features of its environment matter for achieving its goals. Under partial observation, however, different hidden dynamics can produce the same observable experience. This leaves an identification question even for an optimal agent: which environment properties must its choices reveal? Connections between competent decisions and world models have been established under several task interfaces (Richens & Everitt, 2024; Richens et al., 2025). Cifuentes (2026, Section 4.2) explicitly poses the identification question for observation-predicate goals as an open problem. These goals name only observations and actions.

We resolve this open problem on the finite communicating model classes specified below. The answer is an if-and-only-if characterization, with an information converse and constructive recovery. To describe it, imagine an agent with no competence at all: it chooses each action independently and uniformly, and records the resulting observations and actions over a long run. We call their stationary observation–action distribution the *random-walk law*, denoted by J; it is exactly the environment information guaranteed by competence. Because uniform actions assign positive probability to every finite action word, the law retains the controlled predictions associated with observable histories. Hidden labels and redundant hidden states can vary while all these predictions agree.

The characterization has two readings for agent research. A black-box agent that is competent on supplied goals can be turned into a predictive model of its environment: supplying goals and recording actions recovers every controlled prediction that observable histories support, the input to model extraction and to planning with the extracted model. The same characterization bounds what such a probe can guarantee. No decoder recovers a quantity outside J from every competent policy, so competence alone does not guarantee hidden-state labels, redundant states, or transition structure behind equal predictions; a particular policy may still be indiscreet in how it breaks ties between equally good actions.

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.

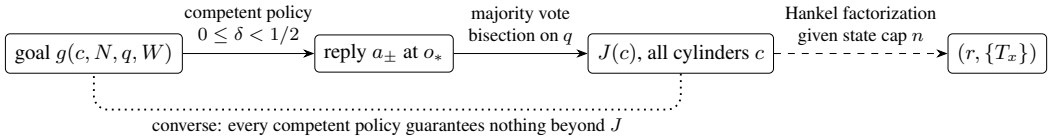

Figure 1: **From native goals to predictive dynamics.** Solid arrows are the recovery of Section 4: sampled root actions estimate the random-walk law J at fixed loss $\delta < 1/2$ under average competence on the queried goals. The dashed arrow also needs a supplied state cap, which permits consistent Hankel dimension and operator estimation. The dotted edge is the converse: competence guarantees nothing beyond J.

The competence premise concerns one policy acting from observation history. For every supplied goal and every hidden start (initial hidden state), its success probability must attain a fixed fraction of the optimum, as in Cifuentes (2026, Definition 3). We call a model admissible if it has such a policy; exact competence requires optimality. If the requirement were only averaged over the stationary initial distribution, models with equal random-walk laws would have the same competent policies and the converse would be immediate; recovery needs only this average.

Theorem A (The information guaranteed by competence, informal)**.** *For admissible finite communicating models with common visible labels and a common possible initial observation, a model quantity is recoverable from every competent policy exactly when it is determined by the random-walk law. This holds at exact competence within any declared admissible subclass, and at fixed relative loss below one half on the full finite admissible class without a common state cap.*

Theorem 3.1, Corollary 3.4, and Theorem 3.6 give the formal information converse: every quantity guaranteed by competence must be determined by J. The quantifier over every policy matters: an individual policy can reveal additional information through its choices between equally good actions. At exact competence, equal laws give identical complete sets of optimal policies. At positive loss those sets may be disjoint. We connect the models through a finite common predictive realization, which forces any universally valid decoder to agree at both endpoints.

Constructive recovery attains this information bound using sampled actions at one supplied observation (Theorem 4.1). We query goals that compare a long-run frequency with a threshold. Averaging removes the effect of the root action on the hidden state, and repeated replies estimate the frequency. Average competence over the queried goal families suffices; no environment trajectory or supplied mixing bound is required for consistency. Figure 1 follows this construction to a predictive model. The separate report bit and known lotteries used by Nayebi (2026, Section 5) are replaced here by environment actions and observation–action goals.

A known state cap changes the converse because the common realization must obey that cap. With deterministic emissions, three states preserve the J characterization; four states can already force an additional bit at positive loss. We explain this distinction through the geometry of predictive mixtures, then evaluate action-based recovery on controlled models with history-dependent predictions.

All of these thresholds are tight: at loss one half, fully observed models with unequal laws can share a competent policy; at positive loss, goals decided within a fixed horizon leave minimax error one half for a stationary state probability (Proposition 4.4), and finite goals admit no uniformly valid finite stopping rule without a mixing bound (Theorem D.6).

2 Models, goals, and observable information

A model $M = (S, \mathcal{A}, P, \mathcal{O}, \Omega)$ has finite hidden states S, finite observations $\mathcal{O}$, and $m = |\mathcal{A}| \ge 2$ actions. At time t, it emits $O_t \sim \Omega(\cdot \mid S_t)$, a policy chooses A_t, and the state transitions according to the row-stochastic matrix P_{A_t}. Emissions and transitions have conditional Markov independence. We assume communication: the union of the actions' positive-transition graphs is strongly connected. Thus $\overline{P} = m^{-1} \sum_a P_a$ is irreducible and has a strictly positive stationary row η_M.

Goals on observations and actions. A *native goal* is a finite disjunction of finite sequences of tests on subsets of $\mathcal{O} \times \mathcal{A}$, as in the observation-predicate language of Cifuentes (2026, Definitions 2–4).

A current test checks the present pair. A next test advances one pair. A first-eventual test waits for the first matching pair, including the present pair; subsequent current tests inspect that same pair. Failure to find a required match means failure of the goal. Success has a finite witness, although eventual waiting can be arbitrarily long. The *depth* is the largest number of tests in a disjunct, and the *width* is the number of disjuncts. Write $\mathcal{G}$ for these goals. Current/next tests express every event on a finite sequence of observation–action pairs.

Competence. A complete behavioral policy $\pi(a \mid h_t, g)$ specifies an action distribution for every goal g and history $h_t = o_0 a_0 \cdots a_{t-1} o_t$. It uses no hidden-state input, and its private randomness is independent of the initial hidden state S_0. For each *hidden start* s, define

$$\begin{aligned} V^{\pi}_{M,s}(g) &= \mathbb{P}^{\pi}_{M,s}(g), \qquad V^{*}_{M,s}(g) = \sup_{\sigma} V^{\sigma}_{M,s}(g), \\ \mathcal{C}_{\delta}(M) &= \{\pi : V^{\pi}_{M,s}(g) \geq (1-\delta) V^{*}_{M,s}(g)\ \forall s, g\}. \end{aligned} \tag{1}$$

The supremum can select a different observation policy for each hidden start; the supplied policy must meet all these benchmarks simultaneously. A model is δ-*admissible* when $\mathcal{C}_{\delta}(M)$ is nonempty. Exact admissibility means $\delta = 0$. The guarantee concerns expected success, including randomness in the policy's actions. Admissibility restricts the class: if two hidden starts share the initial observation and, for some goal, succeed only under different first actions, every policy forgoes at least half of the optimum at one of them, so the model is not δ-admissible for any $\delta < 1/2$. Sufficient conditions compare transition rows of hidden states, or only within an observation group when emissions are deterministic (Lemmas G.17 and G.18).

Access to the policy. Compared models share the visible labels and a supplied possible observation o_*. A policy in $\mathcal{C}_{\delta}(M)$ is also called a competent *oracle*. A *root call* evaluates $\pi(\cdot \mid o_*, g)$ and returns a fresh action, independent conditional on earlier calls. The identifier supplies goals and receives these actions. We are generous to the decoder and stingy to the identifier. The information converses grant the complete policy function, including its values at impossible histories, so that they bound what any decoder could recover: a target F on a declared class of models is *recoverable* when one decoder D satisfies $D(\pi) = F(M)$ for every model M of the class and every $\pi \in \mathcal{C}_{\delta}(M)$. Recovery of J uses root calls only, so that the bound is attained by the weakest access. Decoding the state-cap example of Section 5 additionally evaluates the policy at nine prescribed histories.

The random-walk law. Let $\Sigma = \mathcal{O} \times \mathcal{A}$ and let $\varnothing$ be the empty word. The probability $J_M(w)$ of a pair word $w \in \Sigma^*$ is measured from $S_0 \sim \eta_M$ with independent uniform actions; $J_M(\varnothing) = 1$. For $x = (o, a)$, put $D_o = \mathrm{diag}(\Omega(o \mid s))$ and $B_x = m^{-1} D_o P_a$. Then

$$J_M(x_1 \cdots x_k) = \eta_M B_{x_1} \cdots B_{x_k} \mathbf{1}. \tag{2}$$

For an alternating history h_t ending with an observation, $J_M(h_t)$ is obtained by marginalization. These predictions use the initial distribution η_M; competence is evaluated at the hidden starts in (1).

Admissibility permits predictive information in history. Consider hidden states $0, 1, 2$ with actions a, b and observations x, y. States 0 and 1 emit x and state 2 emits y; action b moves every state to 0; action a moves $0 \to 1$, $1 \to 2$ and $2 \to 0$. Histories xbx and xax both end in observation x, but the next observation under a is x after the first and y after the second, so a predictor has to remember more than the current observation. One policy is nevertheless optimal for every goal from every hidden start: a successful continuation that begins with b works from both x-emitting starts, and when none exists, action a reveals which start occurred (Proposition B.4). The relabeling that exchanges states 0 and 1 shares this policy. The prediction indexed by the hidden label is therefore not recoverable, while the prediction indexed by the observable history is.

3 The complete information in competent behavior

The characterization combines two directions. Different random-walk laws force some goal to elicit different choices. Models with the same law have equal policy values from their respective stationary initial distributions, which controls how much information competence can require a policy to reveal.

3.1 EXACT COMPETENCE

Theorem 3.1 (Complete exact-oracle invariant)**.** *For exactly admissible models M, N with the common interface of Section 2,*

$$J_M = J_N \iff \mathcal{C}_0(M) = \mathcal{C}_0(N) \iff \mathcal{C}_0(M) \cap \mathcal{C}_0(N) \neq \varnothing. \tag{3}$$

The equivalence covers all native goals, including first-eventual tests.

To relate the initial distribution η_M to individual hidden starts, let $\mathcal{C}_0^{\eta}(M)$ be the policies optimal for every goal from the single initial law η_M, with the optimum taken after averaging. Positive stationary weights ensure that a zero average optimality gap has zero gap at every state:

Lemma 3.2 (Positive mixtures detect every exact gap)**.** *For an exactly admissible model, $\mathcal{C}_0(M) = \mathcal{C}_0^{\eta}(M)$. Any strictly positive hidden initial law can replace η_M.*

The other direction uses goals that measure long empirical frequencies. Since the uniformly controlled hidden chain is irreducible, these frequencies concentrate uniformly over hidden starts, including for periodic chains. A threshold between two unequal cylinder probabilities then forces opposite root preferences. This gives

Lemma 3.3 (A common competent policy determines the law)**.** *If two models share a policy satisfying* (1) *for all finite current/next goals at loss $0 \leq \delta < 1/2$, then $J_M = J_N$.*

The finite-chain bound and separating goals are in Appendices B.2 and B.3. They allow both root actions to change the hidden state: concentration applies from either resulting distribution.

Proof of Theorem 3.1. We transfer policy values through the random-walk law. For any fixed policy σ and goal g, the history probability is

$$\mathbb{P}^{\sigma}_{M,\eta_M}(h_t) = m^t J_M(h_t) \prod_{j=0}^{t-1} \sigma(a_j \mid h_j, g). \tag{4}$$

The policy factors depend only on the recorded prefixes, so replacing uniform actions gives this identity, including zero-probability histories. If $J_M = J_N$, every finite transcript has the same probability under σ in the two models. Native success is the increasing union of its finite-time acceptance events. Continuity from below gives equal values for every policy and goal, and thus equal stationary optimizer sets. Lemma 3.2 identifies these with the statewise sets. They are nonempty by admissibility. Lemma 3.3 gives the last implication in

$$J_M = J_N \implies \mathcal{C}_0(M) = \mathcal{C}_0(N) \implies \mathcal{C}_0(M) \cap \mathcal{C}_0(N) \neq \varnothing \implies J_M = J_N. \qquad \square$$

Corollary 3.4 (Exactly which quantities are recoverable)**.** *Let $\mathcal{K}$ be any declared subclass of the exactly admissible models above. For a target F on $\mathcal{K}$, there is a decoder D satisfying $D(\pi) = F(M)$ for every $M \in \mathcal{K}$ and every $\pi \in \mathcal{C}_0(M)$ if and only if*

$$J_M = J_N \implies F(M) = F(N) \quad (M, N \in \mathcal{K}). \tag{5}$$

Equivalently, $F = \tilde{F} \circ J$ for a function on the random-walk laws realized in $\mathcal{K}$. The decoder is an unrestricted function of the complete oracle.

The declared class may fix a state set, a state cap or an emission map. Thus the characterization applies to the information specified by the class as well as to the full model family.

3.2 POSITIVE LOSS AND COMMON PREDICTIVE REALIZATIONS

Positive loss allows different hidden starts to trade off their optimality gaps. Equality of the random-walk laws need not give equal competent-policy sets.

Proposition 3.5 (Nontransitive compatibility)**.** *For every $0 < \delta < 1/2$, equal-J finite communicating δ-admissible models can have disjoint competent-oracle sets. There are three such models for which the first and second each share an oracle with the third but share none with each other.*

Pairwise equality of policy sets no longer supplies the information converse. We connect the models through a common intermediate realization, which we call a *bridge*.

Theorem 3.6 (Positive-loss guaranteed information on the full finite class)**.** *For each fixed $0 \le \delta < 1/2$ on the full finite δ-admissible class, two models have equal J if and only if some finite communicating δ-admissible model R shares a competent oracle with each of them, possibly a different oracle for each pair. A target is recoverable from every complete competent oracle if and only if it is a function of J on this class.*

To construct the bridge, we work with predictions of controlled tests. A controlled test is an alternating observation–action word ending with an observation, or the empty word. Let $f_{M,s}$ give its probability from hidden state s, with actions prescribed and empty-test value one. The hidden-state test functions of two finite models span a finite-dimensional space. In this space define

$$q_{M,s,a} = \sum_{s'} P_a(s, s') f_{M,s'}, \qquad \mathcal{P}_M = \operatorname{conv}\{f_{M,s} : s \in S_M\}. \tag{6}$$

Thus $\mathcal{P}_M$ consists of predictions from hidden-state mixtures, while $q_{M,s,a}$ gives the predictions after action a. The linear prefix operator $\mathcal{L}_{o,a}$ satisfies $(\mathcal{L}_{o,a} f)(w) = f(oaw)$ and $(\mathcal{L}_{o,a} f)(\varnothing) = f(o)$, so $\mathcal{L}_{o,a} f_{M,s} = \Omega_M(o \mid s) q_{M,s,a}$.

Equal J places the common controlled-test function determined by J in $\mathcal{P}_M \cap \mathcal{P}_N$. To transfer competence, we first select endpoint policies that satisfy the relative guarantee conditional on the initial observation. Write $V^{\pi}_{M,s,o}(g)$ for success conditional on $S_0 = s, O_0 = o$, and $V^{*}_{M,s,o}(g)$ for its supremum, whenever $\Omega(o \mid s) > 0$. Prefixing each goal with its initial-observation gate yields such a conditionally competent policy from any statewise competent one (Lemma C.1). The following construction then transfers these policies through common mixtures.

Lemma 3.7 (A finite realization that transfers competence)**.** *If $\mathcal{P}_M \cap \mathcal{P}_N \neq \varnothing$, there is a finite state-emitting model R that realizes every point of this intersection from a suitable hidden initial distribution. Every conditionally competent policy of either endpoint is statewise competent in R.*

Idea of the proof. By the Minkowski–Weyl theorem (Ziegler, 1995, Section 1.1), the intersection is the convex hull of finitely many points. Conditioning a generator on an observation and a prescribed action gives another point of the intersection, together with posterior weights over the hidden states of either endpoint. A generator cannot itself serve as a hidden state, because its successor mixture depends on the observation just emitted. We therefore decompose the successor data of each generator into finitely many extreme tuples and make each tuple one hidden state, whose emission law and action kernels are then fixed. An induction on test length shows that mixtures of these states realize every generator. Competence transfers because, conditional on an emitted observation, the controlled future in R is a mixture of endpoint conditional futures: a conditionally competent endpoint policy attains the required fraction of each term, and averaging over the initial emission gives statewise competence. Appendix C gives the construction and its verification.

The resulting model may have several closed communicating classes. We initialize it to realize the common random-walk law. Concentration of cylinder frequencies forces every class with positive stationary weight to have that same random-walk law. Restricting to one such class preserves its action kernels and the transferred competence, giving the communicating model required by the theorem (Lemma C.2).

A decoder valid for every competent policy must assign the same target value across either shared-policy pair, and hence across M–R–N. Intersections and state splitting can increase the number of hidden states. This is why the theorem allows every finite size, and why a prescribed state cap can change the answer in Section 5.

4 Recovering the random-walk law from actions

We recover a cylinder probability by asking the policy to choose between a frequency event and its complement. The goals prescribe their continuation actions, making the event observable through the initial choice. For a nonempty pair cylinder $c \in \Sigma^d$ and $N \ge 2$ windows, write

$$Z_{N,c} = \frac{1}{N} \sum_{t=1}^{N} \mathbf{1}\{(O_t, A_t), \ldots, (O_{t+d-1}, A_{t+d-1}) = c\}. \tag{7}$$

Fix a known loss $0 \le \delta < 1/2$ and write $\Delta = 1 - 2\delta$ for the available margin and $e_0 = \Delta/[4(2-\delta)]$ for the error probability it tolerates. For a rational cutoff q, let $L = N + d - 1$ and draw a uniform word $W \in \mathcal{A}^L$. A goal $g(c, N, q, W)$ gates on $O_0 = o_*$ and requires W in its continuation. It accepts after root action a_+ when $Z_{N,c} \ge q$, and after a distinct action a_- otherwise.

The event is a native goal. Partition observations into the labels occurring in c and their complement, giving $k_c \le d + 1$ nonempty cells. Each length-L cell word specifies one disjunct and its accepting root marker, so

$$\mathrm{depth}(g) = N + d, \qquad \mathrm{width}(g) = k_c^{N+d-1}. \tag{8}$$

For example, with binary observations, $c = (1, \mathsf{a})$, $N = 3$, $q = 2/3$ and continuation aaa, the triples $011, 101, 110, 111$ require the event marker and the other four require the complement marker. This goal has depth four and width eight. Appendix H.8.1 writes its tests explicitly.

Competence on the queried goals. Recovery only needs success averaged over the prescribed words. Fix an initial law b with positive probability of o_*. For each queried family (c, N, q), put $g_W = g(c, N, q, W)$ and require

$$\mathbb{E}_W[V_b^\pi(g_W) \mid \mathcal{F}] \ge (1 - \delta)\mathbb{E}_W[V_b^*(g_W) \mid \mathcal{F}], \qquad 0 \le \delta < 1/2. \tag{9}$$

Here V_b^π and V_b^* evaluate success under b, $\mathcal{F}$ records earlier calls, and W is fresh uniform conditional on them. Statewise competence implies this condition by averaging over the initial law and the word. It also allows accuracy to vary across words and imposes nothing on unqueried goals.

Theorem 4.1 (Recovery from sampled actions)**.** *Fix $0 \le \delta < 1/2$ and an initial law assigning positive probability to o_*. Under* (9)*, fresh sampled root actions suffice for almost sure, coordinatewise recovery of J. Each stage uses finitely many finite goals, with no supplied state cap or mixing bound. If a state cap n is given, cylinders of length at most $2n - 1$ suffice to consistently recover the minimal predictive dimension and one linear realization.*

Why the comparison works. Averaging over W produces independent uniform continuation actions. The frequency event concentrates on the correct side of $J_M(c)$, uniformly over both post-marker hidden distributions. Lemma B.1 in Appendix B.2 bounds the probability that $Z_{N,c}$ deviates from $J_M(c)$ by more than a tolerance $\gamma > 0$ by $16G_{M,c}^2/(N\gamma^2)$, uniformly over hidden starts. The constant $G_{M,c}$ is the sup norm of a Poisson-equation solution for the finite window chain and needs no aperiodicity. The next two bounds hold under (9) by Proposition D.4.

Lemma 4.2 (Random goals provide a threshold test)**.** *Draw W uniformly from $\mathcal{A}^L$ and query one root action for $g(c, N, q, W)$. If*

$$N \ge \max\left\{2, \frac{64(2-\delta)G_{M,c}^2}{\Delta\gamma^2}\right\}, \tag{10}$$

then the probability of receiving a_+ is at least $1/2 + \Delta/4$ when $J_M(c) \ge q + \gamma$, and at most $1/2 - \Delta/4$ when $J_M(c) \le q - \gamma$. A majority of K independent replies has error at most $\exp(-K\Delta^2/8)$ outside this band, also conditional on earlier adaptive tests.

Theorem 4.3 (Finite accuracy with a model-dependent length)**.** *For $0 < \epsilon < 1/2$ and $0 < \zeta < 1$, choose*

$$B = \left\lceil \log_2 \frac{1}{\epsilon} \right\rceil, \quad K = \left\lceil \frac{8}{\Delta^2} \log \frac{B}{\zeta} \right\rceil, \quad N \ge \max\left\{2, \frac{256(2-\delta)G_{M,c}^2}{\Delta\epsilon^2}\right\}. \tag{11}$$

B rounds of ordinary bisection on $[0, 1]$, using the sampled majority test at each midpoint, return an estimate with $\mathbb{P}(|\widehat{J}(c) - J_M(c)| \le \epsilon) \ge 1 - \zeta$. The procedure uses BK root replies and the goal sizes in (8)*.*

Increasing the lengths with summable comparison errors gives the pointwise schedule of Theorem D.1. With Proposition D.4, this schedule proves the recovery claim of Theorem 4.1; its realization claim is Corollary F.1. The threshold and finite-accuracy proofs are in Appendices D.1 and D.2.

The representation step is classical: Hankel factorization turns these cylinder probabilities into a minimal linear prediction system (Littman et al., 2001; Jaeger, 2000; Thon & Jaeger, 2015). Its state is indexed by observable history. Corollary F.1 gives the operators and their consistent estimates; identifying hidden parameters additionally needs the rank conditions in Corollary F.3, which require at least as many observations as hidden states and so exclude the aliased models of Section 6.

Cost of a query. The identifier pays in root replies: Theorem 4.3 uses BK replies per cylinder. The width in (8) counts all length-L cell words, one disjunct each, so this encoding grows exponentially with the horizon. The goal is evaluated by a finite count monitor that stores the matched prefix and a capped occurrence count and accepts exactly the paths of the native disjunction. With known model parameters its arithmetic is polynomial in the numerical horizon and count threshold (Proposition H.3). A responding planner can compute values in this representation and expose only its chosen actions; its model and value computation determine the competence guarantee (Appendix H.2).

A supplied hitting-time bound makes the required length explicit through $G_{M,c} \leq d + m^{-d}\bar{\tau}_{\mathrm{unif}}$ (Proposition H.2). Communication alone gives the following limits.

Proposition 4.4 (A sharp endpoint and a physical-horizon obstruction)**.** *(i) Two fully observed two-state communicating models with unequal J can share a policy that is statewise $1/2$-competent for every native goal. (ii) For every physical horizon $H \geq 1$ and $0 < \delta < 1/2$, the minimax expected absolute error in recovering a stationary binary-state probability from behavior is exactly $1/2$ over a two-state communicating class, when queried goals must be decided within H transitions. This holds even if complete policy functions for the permitted goals are supplied and each oracle is competent at every native depth.*

Appendix E proves both parts. At positive loss, root access restricted to finite current/next goals also admits no uniformly valid, almost surely finite stopping certificate under a fixed initial law, even with two states (Appendix D.4.3).

5 A STATE CAP CAN FORCE ADDITIONAL INFORMATION

A state cap constrains which models can connect two equal-law realizations. Within the declared class, join two models when they share a competent policy. A *compatibility component* contains the models connected by a finite chain of these edges. Every recoverable target is constant on each component, because each shared policy must receive the same decoded value. Theorem 3.6, whose bridges may have any finite size, makes each fiber of the random-walk law one component; a state cap can split it.

Theorem 5.1 (The smallest deterministic-emission state cap)**.** *For $0 \leq \delta < 1/2$, recoverable targets are exactly functions of J on the full admissible class with deterministic emissions and at most three hidden states. For every $29/100 \leq \delta < \delta_*$, where $\delta_* > 293/1000$, an explicit random-walk law J_0 has exactly two compatibility components among all admissible realizations with at most four states, including unknown emissions. On the promised law $J = J_0$, one native goal and sampled actions at nine prescribed histories identify the component with arbitrarily high confidence. Thus four is the smallest deterministic-emission cap allowing information beyond J at some positive loss.*

The loss interval is where the construction is verified; neither endpoint is claimed to be a critical loss. In the exactly admissible class every nonempty equal-law fiber is one component by Theorem 3.1, so the additional bit is a positive-loss phenomenon. Both orientations are shown admissible for every native goal from $\delta = 29/100$ on, and the decoding gap Γ_δ between the components' possible mean scores decreases linearly to zero at δ_*.

The geometric distinction. With three deterministic-emission states and at least two possible observations, each repeated-observation group has at most two states, so its predictive hull is a point or segment. Intersections preserve this geometry and admit a bridge within the cap. If only one observation can occur under deterministic emissions, goals depend only on actions. Four states allow three states to share one observation and form a triangle of predictive mixtures. The two orientations in Figure 2 have the same random-walk law, but their triangular hulls intersect in a hexagon. Realizing that intersection directly uses seven states, including the state emitting the other observation.

The proof also classifies every admissible realization of J_0 under the four-state cap. It rules out alternative four-state bridges and shows that competent branch choices separate the two orientations throughout the given loss interval. This classification supplies the additional bit. Theorem G.1 gives the full statement and sampling bound; Appendix G develops the realization geometry.

Reading the bit from actions. The distinguishing goal has three successful action words sharing a prefix tree. Their nine nonterminal prefixes give the history inputs. We follow sampled actions

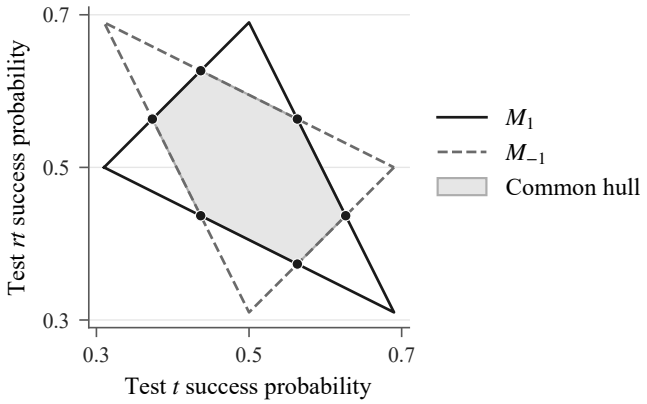

Figure 2: **Why the state cap matters.** The two endpoint realizations M_1, M_{-1} have triangular predictive hulls for their three zero-emitting states. Their intersection has six vertices, so the intersection construction uses seven states including the one-emitting state. The full realization classification separately excludes every alternative four-state bridge.

through this tree and assign a fixed score to a completed word, using at most five calls per replicate. Competence places the two components' mean scores in disjoint intervals, so their midpoint provides a decision threshold. Appendix G.4 specifies the words, weights and confidence bound. Corollary G.16 in Appendix G.6 extends the three-state case to stochastic emissions under rank conditions.

6 Illustrating recovery from actions

We illustrate recovery with planners computed from finite models. The known model fixes the planner's competence and the target predictions; the identifier receives only sampled root actions and the public query parameters.

In the three-state memory example, recovery at length $N = 512$ gives next-observation MAE 0.00782 at exact competence and 0.01999 at loss 0.25, compared with 0.19028 for a predictor using only the current observation. All ten runs, five seeds at each loss, select predictive rank three. Appendix I.3 reports the full 120-run grid, including IID, fully observed and relabeled models.

Table 1 and Figure 3 consider three models with stochastic transitions and aliased observations. We estimate 42 cylinders through length three from sampled root actions and form a calibrated two-pair-context predictor with count-based backoff. Three competent policies at loss 0.35 answer the queries: least-left and greatest-left (L, G) take the two endpoints of the admissible root-action interval, and event-adverse (A) takes the endpoint that minimizes the queried upper-tail event; each has three paired seeds. The two-pair predictor has lower MAE than the stationary current-pair predictor in 27 of 27 fits, than the best current-pair predictor in 26 of 27, and than a trajectory-based Hankel/PSR learner given the same number of environment transitions on all three models. The last comparison changes the estimator as well as the access: the same two-pair context predictor applied to trajectory frequencies at 10^6 transitions gives MAE 0.00130–0.00157 (Appendix I.6), so the gap in Table 1 reflects the estimator, and root replies match passive sampling once it is held fixed. Appendix I.5 specifies the estimator, word reuse and query menu, and Appendix I.7 the trajectory learner.

Each stochastic-model fit uses 1,002,300 root replies; its goals have depth 2051 and native width 2^{2050}, so nobody writes them out by hand and count monitors evaluate them (Appendix I.5 reports costs and per-seed results). Further illustrations include component ablations (Appendix I.6) and all 18 fits on 6-, 8- and 12-state models with an exact planner under a fixed initial law (Appendix I.8).

7 Related work and conclusion

Information in behavior. Competence-to-model results depend on the task and response interface (Richens & Everitt, 2024; Richens et al., 2025; Cifuentes, 2026). Standard fixed-dynamics inverse RL

Table 1: **Next-observation MAE under stochastic aliasing.** All positive-probability histories of five and six pairs. Root columns average three competent policies and three paired seeds at loss 0.35. The trajectory PSR learner uses as many environment transitions as a root fit uses replies; its column gives mean and sample SD over three seeds. Current-pair columns are population predictors computed from the model.

Model	Current pair, J	Best current pair	Trajectory PSR (SD)	One-pair root	Two-pair root
Three-state	.00623	.00617	.00377 (.00075)	.00666	.00135
Four-state	.00288	.00258	.00412 (.00153)	.00329	.00119
Four-state draw	.00273	.00244	.00732 (.00538)	.00310	.00123

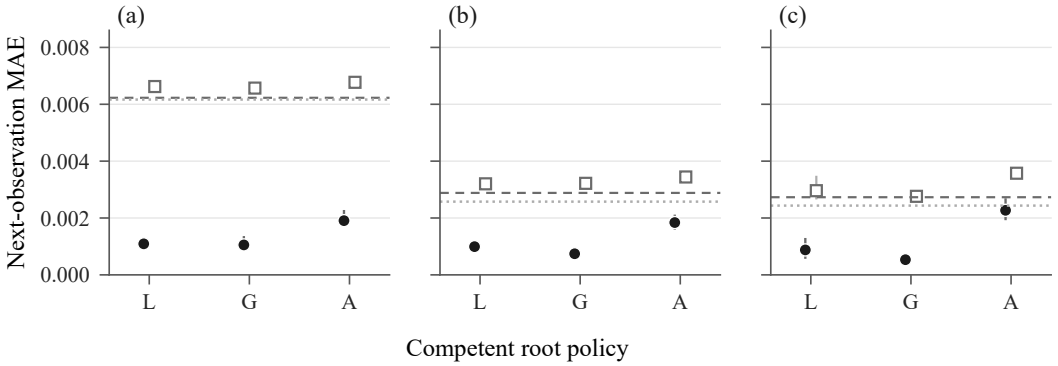

Figure 3: **Behavioral forecasting under stochastic aliasing.** (a) Three-state model; (b) four-state model; (c) four-state draw. Black circles use two-pair contexts with backoff; white squares use one-pair contexts from the same acquisition and calibration. Bars span three paired seeds. Dashed lines give stationary current-observation predictions; dotted lines give the best current-observation predictor under absolute-error loss. L/G/A are the three competent policies defined in the text.

identifies rewards up to ambiguities in observed behavior (Ng & Russell, 2000; Skalse et al., 2023); here the supplied goals are known and the environment is unknown. The closest predictive recovery result, Nayebi (2026, Section 5), uses a separate report bit and known lotteries, with consistency as average regret vanishes. Our frequency construction handles environment-changing root actions at fixed relative loss below one half, and the information converse characterizes every recoverable target. At exact competence the equivalence follows the classical intuition that predictive-state representations capture all observable dynamics (Littman et al., 2001). At positive loss it does not: equal laws can have disjoint competent-policy sets (Proposition 3.5), the converse needs the common-realization construction, and a state cap can break it (Theorem 5.1). Appendix H compares these interfaces with action equivalence and numerical value queries (Lu et al., 2026; Batra, 2026; Letcher et al., 2026).

Predictive representations. Spectral HMM/PSR learners estimate the moments of PSRs and observable operator models (Littman et al., 2001; Jaeger, 2000; Thon & Jaeger, 2015) from trajectories (Hsu et al., 2012; Boots et al., 2011); we obtain them through goal-conditioned actions. A single trajectory supplies dependent observations, while fresh replies require a competent policy to answer them; Appendix H.6 compares these sampling assumptions. Hidden-parameter identification uses additional latent-class conditions (Allman et al., 2009). Value equivalence and predictive channels are discussed in Appendix H (Grimm et al., 2020; Rosas et al., 2025; Baltieri et al., 2026).

A competent agent cannot help showing us what a random walk sees, and it owes us nothing more. The common-realization construction carries this to positive loss; a known state cap is the exception, where four states can force one more bit.

AI DISCLOSURE

Generative AI tools, including Codex and Claude, assisted with mathematical exploration, literature searches, proof checking, and prose polishing. In particular, they were used to explore the averaging and separation arguments underlying Theorem 3.1, especially Lemmas B.1, 3.3 and 3.2 and Corollary 3.4, and the root-action extraction underlying Theorems 4.3 and D.1, and to search for counterexamples to intermediate claims with linear programs and SMT solvers. They were not used to generate data sets directly; all data in the numerical examples come from simulations of synthetic finite models. The authors conceived the paper's core ideas, developed its methods, and verified the correctness and originality of all content, including the references.

REPRODUCIBILITY STATEMENT

All assumptions on the model class, the goal language, and the competence interface are stated in Section 2, and complete proofs are given in the main text and appendix. Appendices I, I.4 and I.5 describe the models, estimators, sampling schedules, paired repetitions and query costs used in Section 6. The source code, synthetic data, recorded experimental outputs, and finite exact certificates are available at `https://github.com/submissionrepo/world-model-psr`, with instructions for replaying the reported results and rerunning the experiments.

REFERENCES

Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009. doi: 10.1214/09-aos689. URL https://doi.org/10.1214/09-aos689.

Brian D. O. Anderson. The realization problem for hidden Markov models. *Mathematics of Control, Signals and Systems*, 12:80–120, 1999. doi: 10.1007/PL00009846. URL https://link.springer.com/article/10.1007/PL00009846.

Manuel Baltieri, Filippo Torresan, Yivan Zhang, Alexander Boyd, and Fernando E. Rosas. World models of environment, agent and joint agent-environment systems. arXiv preprint arXiv:2608.20401, 2026. URL https://arxiv.org/abs/2608.20401v1. Version 1.

Neal Batra. From optimal actions to world models: Identifiability of transition kernels in discounted MDPs, 2026. URL https://arxiv.org/abs/2608.07301v1. Version 1.

David Blackwell and Lambert Koopmans. On the identifiability problem for functions of finite Markov chains. *The Annals of Mathematical Statistics*, 28(4):1011–1015, 1957. doi: 10.1214/aoms/1177706802. URL https://doi.org/10.1214/aoms/1177706802.

Byron Boots, Sajid M Siddiqi, and Geoffrey J Gordon. Closing the learning-planning loop with predictive state representations. *The International Journal of Robotics Research*, 30(7): 954–966, 2011. doi: 10.1177/0278364911404092. URL https://doi.org/10.1177/0278364911404092.

Santiago Cifuentes. General agents contain world models, even under partial observability and stochasticity, 2026. URL https://arxiv.org/abs/2602.03146v1. Version 1.

Xiang Gao, Kaiwen Dong, Yuguang Yao, Padmaja Jonnalagedda, and Kamalika Das. Textual belief states for world models: Identifiable representation learning under strict mediation. arXiv preprint arXiv:2606.27681, 2026. URL https://arxiv.org/abs/2606.27681v1. Version 1.

Peter W. Glynn and Alex Infanger. Solution representations for Poisson's equation, martingale structure, and the Markov chain central limit theorem. *Stochastic Systems*, 14(1):47–68, 2023. doi: 10.1287/stsy.2022.0001. URL https://pubsonline.informs.org/doi/10.1287/stsy.2022.0001.

Christopher Grimm, Andre 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://papers.nips.cc/paper/2020/hash/3bb585ea00014b0e3ebe4c6dd165a358-Abstract.html.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. A spectral algorithm for learning Hidden Markov Models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012. doi: 10.1016/j.jcss.2011.12.025. URL https://doi.org/10.1016/j.jcss.2011.12.025.

Qingqing Huang, Rong Ge, Sham Kakade, and Munther Dahleh. Minimal realization problems for hidden Markov models. *IEEE Transactions on Signal Processing*, 64(7):1896–1904, 2016. doi: 10.1109/TSP.2015.2510969. URL https://arxiv.org/abs/1411.3698v2.

Hisashi Ito, Shun-Ichi Amari, and Kingo Kobayashi. Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Transactions on Information Theory*, 38(2): 324–333, 1992. doi: 10.1109/18.119690. URL https://doi.org/10.1109/18.119690.

Herbert Jaeger. Observable operator models for discrete stochastic time series. *Neural Computation*, 12(6):1371–1398, 2000. doi: 10.1162/089976600300015411. URL https://doi.org/10.1162/089976600300015411.

Stefan Kiefer, Andrzej S. Murawski, Joël Ouaknine, Björn Wachter, and James Worrell. On the complexity of the equivalence problem for probabilistic automata. In Lars Birkedal (ed.), *Foundations of Software Science and Computational Structures (FoSSaCS)*, volume 7213 of *Lecture Notes in Computer Science*, pp. 467–481. Springer, 2012. doi: 10.1007/978-3-642-28729-9_31. URL https://arxiv.org/abs/1112.4644v2.

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, 2026. URL https://arxiv.org/abs/2606.21173v1. Version 1.

Michael L. Littman, Richard S. Sutton, and Satinder Singh. Predictive representations of state. In *Advances in Neural Information Processing Systems*, volume 14, pp. 1555–1561. MIT Press, 2001. URL https://proceedings.neurips.cc/paper_files/paper/2001/file/1e4d36177d71bbb3558e43af9577d70e-Paper.pdf.

Yikai Lu, Yifei Wu, Xinyu Lu, and Tongxin Li. World models in pieces: Structural certification for general agents. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL https://arxiv.org/abs/2606.24842v1.

Aran Nayebi. What capable agents must know: Selection theorems for robust decision-making under uncertainty. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 4773–4795. PMLR, 2026. URL https://proceedings.mlr.press/v337/nayebi26a.html.

Andrew Y. Ng and Stuart Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 663–670, 2000. URL https://ai.stanford.edu/~ang/papers/icml00-irl.pdf.

Henri Poincaré. *Science and Hypothesis*. Walter Scott Publishing, London, 1905. URL https://brocku.ca/MeadProject/Poincare/Poincare_1905_01.html.

Miquel Ramírez and Hector Geffner. Plan recognition as planning. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence*, pp. 1778–1783, 2009. URL https://www.ijcai.org/Proceedings/09/Papers/296.pdf.

Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *International Conference on Learning Representations*, volume 2024, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/44a2b9f7bf9aec3f1fa333ad875b0ee0-Paper-Conference.pdf.

Jonathan Richens, Tom Everitt, and David Abel. General agents need world models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 51659–51687. PMLR, 2025. URL https://proceedings.mlr.press/v267/richens25a.html.

Fernando E. Rosas, Alexander Boyd, and Manuel Baltieri. AI in a vat: Fundamental limits of efficient world modelling for agent sandboxing and interpretability. *Reinforcement Learning Journal*, 6:2844–2881, 2025. URL https://rlj.cs.umass.edu/2025/papers/Paper367.html.

Mohamad Louai Shehab, Antoine Aspeel, and Necmiye Ozay. Active reward machine inference from raw state trajectories. arXiv preprint arXiv:2604.07480, 2026. URL https://arxiv.org/abs/2604.07480v1. Version 1.

Joar Max Viktor Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, and Adam Gleave. Invariance in policy optimisation and partial identifiability in reward learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32033–32058. PMLR, 2023. URL https://proceedings.mlr.press/v202/skalse23a.html.

Kaizhen Tan, Xin Xu, Siru Tao, Yixiao Li, Hanzhe Hong, Yang Feng, and Heqing Du. What can latent world models know? physical parameter identifiability in multimodal predictive representations. arXiv preprint arXiv:2607.27017, 2026. URL https://arxiv.org/abs/2607.27017v3. Version 3.

Michael Thon and Herbert Jaeger. Links between multiplicity automata, observable operator models and predictive state representations – a unified learning framework. *Journal of Machine Learning Research*, 16(4):103–147, 2015. URL https://jmlr.org/papers/v16/thon15a.html.

Wen-Guey Tzeng. A polynomial-time algorithm for the equivalence of probabilistic automata. *SIAM Journal on Computing*, 21(2):216–227, 1992. doi: 10.1137/0221017. URL https://epubs.siam.org/doi/10.1137/0221017.

Roi Weiss and Boaz Nadler. Learning parametric-output HMMs with two aliased states. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 635–644. PMLR, 2015. URL https://proceedings.mlr.press/v37/weiss15.html.

Xiangteng Zhang, Yang Guan, Bo Zhang, Hongyang Li, Ya-Qin Zhang, and Shengbo Eben Li. On the identifiability of controlled world models. arXiv preprint arXiv:2607.22430, 2026. URL https://arxiv.org/abs/2607.22430v2. Version 2.

Günter M. Ziegler. *Lectures on Polytopes*, volume 152 of *Graduate Texts in Mathematics*. Springer, New York, 1995. doi: 10.1007/978-1-4613-8431-1. URL https://link.springer.com/book/10.1007/978-1-4613-8431-1.

A NOTATION

Symbols used in more than one section are listed with the place where they are defined. Symbols declared inside a single proof, and the constants of the four-state construction in Appendix G, are introduced where they are used. Concrete actions and observations in examples are set in sans-serif, such as a and x.

Symbol	Meaning	Defined in
Models, goals and competence		
M, N	compared models	Section 2
$S, \mathcal{O}, \mathcal{A}; m$	hidden states, observations, actions; $m = \lvert\mathcal{A}\rvert$	Section 2
$P_a, \Omega; \overline{P}$	transition matrices and emission law; their uniform-action average	Section 2
S_t, O_t, A_t	hidden state, observation and action at time t	Section 2
η_M	stationary row of $\overline{P}$	Section 2
$g; \mathcal{G}$	native goal; the set of native goals	Section 2
$\pi, \sigma; h_t$	policies; a history ending with O_t	Section 2
$V^{\pi}_{M,s}(g), V^{*}_{M,s}(g)$	success probability from hidden start s, and its optimum	(1)
$V^{\pi}_{M,s,o}(g), V^{*}_{M,s,o}(g)$	the same, conditional on $O_0 = o$	Section 3
$\delta; \mathcal{C}_\delta(M)$	relative loss; competent policies (oracles)	(1)
o_*	supplied root observation	Section 2
$F; D$	target; decoder	Section 2
$\tilde{F}$	function of the random-walk law with $F = \tilde{F} \circ J$	Corollary 3.4
R	a bridge between two models	Section 3
Random-walk law and predictions		
$\Sigma; \varnothing$	pair alphabet $\mathcal{O} \times \mathcal{A}$; empty word	Section 2
J_M, J	random-walk law	(2)
$D_o; B_x$	emission diagonal; pair matrix $m^{-1} D_o P_a$ for $x = (o, a)$	(2)
$f_{M,s}; q_{M,s,a}$	controlled-test predictions from s; the same after action a	(6)
$\mathcal{P}_M; \mathcal{L}_{o,a}$	predictive polytope; prefix operator	(6)
$\mathsf{H}, \mathsf{H}_x; r$	Hankel matrix and its insertions; its rank	Appendix F
α, β, T_x	minimal linear realization	(27)
Recovery from actions		
$c, d; k_c$	pair cylinder and its length; number of observation cells	Section 4
$N; Z_{N,c}$	number of counted windows; empirical frequency	(7)
$q; L; W$	cutoff; word length $N + d - 1$; random continuation word	Section 4
$g(c, N, q, W), g_W$	threshold query	Section 4
$\Delta; e_0$	margin $1 - 2\delta$; tolerated error $\Delta/[4(2 - \delta)]$	Section 4
$b; \mathcal{F}$	initial law; earlier calls	(9)
$G_{M,c}; \gamma$	Poisson constant; comparison tolerance	Lemma B.1
$B; K$	bisection rounds; replies per comparison	Theorem 4.3
$\epsilon; \zeta$	accuracy; failure probability	Theorem 4.3
n	state cap	Theorem 4.1
$\bar{\tau}_\mu$	largest expected hitting time under reference actions μ	Appendix H.7
State cap		
$J_0; M_\tau$	the four-state law; its realizations	Appendix G
$\delta_*; \Gamma_\delta$	end of the loss interval; decoding gap	(65), (66)
$\omega; C$	separating score vector; largest leaf score	(66)

B PROOFS OF THE EXACT INFORMATION CHARACTERIZATION

B.1 PROOF OF LEMMA 3.2

Proof. We fix a goal and an exact statewise optimizer π_0. Every policy has stationary value at most $\sum_s \eta_M(s)V^*_{M,s}(g)$, and π_0 attains this sum. Since the gaps are nonnegative and the weights strictly positive, stationary optimality is equivalent to

$$0 = \sum_s \eta_M(s)\big(V^*_{M,s}(g) - V^\pi_{M,s}(g)\big) \quad \Longleftrightarrow \quad V^\pi_{M,s}(g) = V^*_{M,s}(g) \quad \forall s. \qquad \square$$

B.2 FINITE-CHAIN AVERAGING

Lemma 3.3 and the recovery analysis of Appendix D both use the following concentration bound. It also applies to countable discrete observations, by partitioning the alphabet into the labels in the queried cylinder and their complement.

Lemma B.1 (Uniform averaging over hidden starts)**.** *For each model M and cylinder c, there is a finite constant $G_{M,c}$ such that, for $N \geq 2$ and $\gamma > 0$,*

$$\sup_b \mathbb{P}^{\mathrm{unif}}_{M,b}\big(|Z_{N,c} - J_M(c)| \geq \gamma\big) \leq \frac{16G^2_{M,c}}{N\gamma^2}. \tag{12}$$

Here b is any hidden-state distribution at the beginning of the measured segment. Periodicity is allowed.

Proof. We first describe the chain on which the Poisson equation is solved. Let C_t be the cell containing O_t in the finite cylinder partition. Under uniform actions, $X_t = (S_t, C_t, A_t)$ is a finite chain on triples with positive emission probability, with transition matrix

$$Q_X((s, C, a), (s', C', a')) = P_a(s, s')\Omega(C' \mid s')/m.$$

From any triple we can take one possible successor and then a positive controlled path to any desired hidden state. At its endpoint every feasible observation cell and action have positive probability. The chain is therefore irreducible, with stationary weights $\eta_M(s)\Omega(C \mid s)/m$. Its feasible length-$d$ window chain is also irreducible: connect the last triple of one window to the first of another, then append the rest of the target window. Its stationary cylinder indicator has mean $J_M(c)$.

We recall why the required Poisson solution exists without an aperiodicity assumption. For a finite irreducible transition matrix Q, a limit point of Cesàro averages of probability rows is stationary. Stationarity propagates a positive coordinate along every positive path, so the stationary row ν is strictly positive. A real vector satisfying $Qu = u$ is constant: at a maximizer, every positive successor must also maximize, and irreducibility reaches all states. Thus $I - Q$ has a one-dimensional kernel, and its image is the hyperplane orthogonal to ν. For the window indicator f, we may choose

$$(I - Q)u = f - J_M(c)\mathbf{1}, \qquad \nu u = 0.$$

This fixes the finite norm $G_{M,c} = \|u\|_\infty$. We write Y_t for the window chain and set $D_{t+1} = u(Y_{t+1}) - (Qu)(Y_t)$. These are martingale differences bounded by $2G_{M,c}$, orthogonal at different times. The telescoping identity is

$$\sum_{t=0}^{N-1} \big(f(Y_t) - J_M(c)\big) = u(Y_0) - u(Y_N) + \sum_{t=0}^{N-1} D_{t+1}.$$

The boundary term has absolute value at most $2G_{M,c}$. Squaring, using $(a+b)^2 \leq 2a^2 + 2b^2$, and averaging gives the same bound under every initial window law. Every hidden initial distribution induces such a law. Markov's inequality for the square gives

$$\sup_b \mathbb{P}^{\mathrm{unif}}_{M,b}\big(|Z_{N,c} - J_M(c)| \geq \gamma\big) \leq \frac{8G^2_{M,c}}{\gamma^2}(N^{-2} + N^{-1}) \leq \frac{16G^2_{M,c}}{N\gamma^2}. \qquad \square$$

B.3 Proof of Lemma 3.3

Proof. We turn unequal cylinder probabilities into opposite root choices. We suppose, for contradiction, that the laws differ and exchange model names so that $J_M(c) > J_N(c)$. We choose a rational threshold strictly between them. Lemma B.1 implies that the empirical frequency exceeds this threshold with probability approaching one in M and zero in N, uniformly over their hidden starts, under uniform actions. There are finitely many starts. Averaging the sum of their error probabilities over the action word supplies a single prescribed word for which every error is at most e_0.

We use distinct root actions a_+, a_- and a goal gated by $O_0 = o_*$. The goal requires the prescribed continuation word and accepts the frequency event after a_+ or its complement after a_-. A finite disjunction over observation words expresses it. The root markers only change the hidden distribution at the beginning of the measured segment, so the uniform bounds still apply. Following the prescribed word attains each branch probability; deviations can only remove successes.

The error level e_0 of Section 4 makes the required root probabilities conflict. At a hidden start that can emit o_*, competence for the gated goal cancels its positive emission probability. In M, the a_+ branch has value at least $1 - e_0$ and the other at most e_0. If x is the common root probability of a_+, its total success is at most $x + e_0$. In N the corresponding bound applies to the probability y of a_-. Hence

$$x, y \geq (1 - \delta)(1 - e_0) - e_0 = \frac{1}{2} + \frac{\Delta}{4}, \qquad 1 \geq x + y \geq 1 + \frac{\Delta}{2} > 1. \qquad \square$$

B.4 Proof of Corollary 3.4

Proof. We first suppose a decoder exists. Equal laws give a common exact policy by Theorem 3.1, so the two target values must agree. Conversely, every admitted policy determines a unique law by that theorem. We assign it the target value of its law's fiber, which is well defined by (5). For every admissible pair,

$$D(\pi) = \tilde{F}(J_M) = F(M). \qquad \square$$

B.5 Nonidentification of hidden realizations

We first make the initial-emission example precise. It distinguishes a hidden-start family from the random-walk law identified by the oracle. The two models have action set $\{\mathsf{a}, \mathsf{b}\}$ and observation set $\{\mathsf{x}, \mathsf{y}\}$. In the two-state model M,

$$P_{\mathsf{a}}(i, j) = P_{\mathsf{b}}(i, j) = \tfrac{1}{2} \quad (i, j \in \{1, 2\}), \qquad \Omega_M(\mathsf{x} \mid 1) = \tfrac{9}{10}, \quad \Omega_M(\mathsf{x} \mid 2) = \tfrac{1}{10}. \tag{13}$$

The one-state model N emits either symbol with probability $1/2$. The hidden-start prediction family includes the probabilities of O_0, before conditioning on its realized value.

Lemma B.2 (Finite acceptance and reachability)**.** *Every native goal has a finite deterministic monitor with an absorbing accepting state. In a finite fully observed controlled Markov process with an absorbing accepting set, one deterministic stationary policy maximizes eventual acceptance from every starting state.*

Proof. We record each disjunct's current test position and whether it is waiting for a next pair or a first match. A current mismatch rejects that disjunct. A next test creates an obligation at the following pair. A first-eventual test waits until its first match, then consumes that test at the matching pair. Following current or matching eventual tests process the same pair. Each consumption advances the finite sequence, so processing terminates. Taking the finite product over disjuncts, with acceptance when any component accepts, gives the monitor.

For reachability, we use discounted hitting values and select one optimizer recurring as the discount tends to one. Let $\mathcal{Z}$ be the finite controlled state set, $\mathcal{B}$ the accepting set, and τ its first hitting time. For $0 < \rho < 1$, the operator

$$(\mathcal{T}_\rho v)(z) = \begin{cases} 1, & z \in \mathcal{B}, \\ \rho \max_a \sum_{z'} Q_a(z, z') v(z'), & z \notin \mathcal{B} \end{cases}$$

is a sup-norm contraction. Its fixed point is the maximum of $\mathbb{E}_z^\sigma[\rho^\tau \mathbf{1}_{\{\tau < \infty\}}]$ over arbitrary policies, as follows by iterating finite-horizon hitting values. Choosing a maximizing action at each state

attains the fixed point with a deterministic stationary policy. There are finitely many such policies. We choose a sequence $\rho \uparrow 1$ along which the maximizing policy is a single policy κ. For every starting state and competing policy, monotone convergence gives

$$\mathbb{P}_z^{\sigma}(\tau < \infty) = \lim_{\rho\uparrow 1} \mathbb{E}_z^{\sigma}[\rho^{\tau}\mathbf{1}_{\{\tau<\infty\}}] \le \lim_{\rho\uparrow 1} \mathbb{E}_z^{\kappa}[\rho^{\tau}\mathbf{1}_{\{\tau<\infty\}}] = \mathbb{P}_z^{\kappa}(\tau < \infty). \qquad \square$$

Proposition B.3 (An exact common oracle with different hidden-start families)**.** *The models in* (13) *share an exact complete statewise oracle for every native goal. Their random-walk laws agree, but their hidden-start initial-emission families differ.*

Proof. We optimize after each initial observation, because the remaining controlled process is common. In M, every transition draws a fresh uniform hidden state. Conditional on any hidden start and either initial observation, the later observations are independent fair draws, unaffected by the actions. Model N has the same conditional future law.

For a goal g, we combine its monitor with the current observation. This is a finite fully observed controlled process common to both models. Lemma B.2 gives a policy that attains the optimum $v_g(o)$ from each initial observation o. Every policy selected for any hidden start has conditional value at most $v_g(o)$. Our common policy attains both conditional optima, so it is statewise exact in either model:

$$V_{i,s}^{\pi}(g) = \sum_{o} \Omega_i(o \mid s) v_g(o) = V_{i,s}^{*}(g) \qquad (i \in \{M, N\}).$$

We choose this policy for each goal and common arbitrary actions at irrelevant histories, yielding one complete oracle. Under stationary initialization both models generate the fair IID observation process, while

$$\{\mathbb{P}_{M,s}(O_0 = \mathsf{x}) : s \in S_M\} = \{9/10, 1/10\} \neq \{1/2\} = \{\mathbb{P}_{N,s}(O_0 = \mathsf{x}) : s \in S_N\}. \qquad \square$$

Different hidden kernels with the same controlled process. The numerical controlled model and its four-state realization provide a second illustration. For an observed two-state kernel P_a, let a hidden state be (s, u) with $u \in \{0, 1\}$, emit s, and transition according to

$$\widetilde{P}_a((s,u),(t,v)) = P_a(s,t) W_a(u,v), \tag{14}$$

where W_a is row-stochastic. Summing over v leaves $P_a(s,t)$ for every copy u. The controlled observation process from either copy of s is therefore the same as from observed state s. A goal monitor combined with s gives simultaneous exact optimal policies by Lemma B.2. The stationary marginal on s is the unique stationary distribution of the averaged two-state kernel, so the two models have equal J. Their hidden transition matrices have different dimensions, while their minimal predictive dimensions agree. Appendix I specifies the positive matrices used in the experiment.

B.6 An exactly admissible model whose observations need memory

The three-state example in Section 2 has the following deterministic transitions and observations.

State	Observation	Action a	Action b
0	x	1	0
1	x	2	0
2	y	0	0

Proposition B.4 (History resolves ambiguity, but the labeled root is not identified)**.** *The three-state model in Section 2 is exactly admissible. Let N be its relabeling that exchanges hidden states 0 and 1 and fixes state 2. The two models share an exact complete oracle for all native goals, but*

$$\mathbb{P}_M(O_1 = \mathsf{y} \mid S_0 = 0, \mathrm{do}(A_0 = \mathsf{a})) = 0, \qquad \mathbb{P}_N(O_1 = \mathsf{y} \mid S_0 = 0, \mathrm{do}(A_0 = \mathsf{a})) = 1.$$

The minimax expected absolute error for this target from the common oracle is $1/2$.

Proof. We construct an optimal policy by taking a common successful continuation when one exists and otherwise revealing the initial ambiguity. We fix a native goal g. In a deterministic model, a

reachable success has a finite witness and can be attained with probability one by prescribing a satisfying action sequence.

An initial y identifies state 2, so we choose a satisfying sequence when one exists. From initial x, states 0 and 1 are possible. If a satisfying b-prefixed trajectory exists, we choose it: both starts produce prefix xbx and reach state 0. If none exists, we choose a. The prefixes xax and xay identify the two initial cases and their new states 1 and 2. We choose a satisfying continuation separately in each case where one exists. Otherwise both first-action branches are impossible from that start, whose optimal value is zero. After the first transition the state is known and can be tracked. These disjoint histories define one observation policy, with

$$V^{\pi}_{M,s}(g) = V^{*}_{M,s}(g) \qquad (s \in \{0,1,2\}).$$

We choose such a policy for each goal. Let ι exchange 0 and 1. Its observation-preserving conjugacy makes the observable law under every policy from N, s equal that from $M, \iota(s)$; the same complete oracle is therefore statewise exact in N.

The transition table gives the two target values. Every query transcript, and hence the estimator output $\widehat{p}$, has the same distribution under the common oracle. The constant $1/2$ attains the lower bound

$$\max\{\mathbb{E}|\widehat{p}|, \mathbb{E}|\widehat{p}-1|\} \geq \tfrac{1}{2}\mathbb{E}(|\widehat{p}| + |\widehat{p}-1|) \geq \tfrac{1}{2}. \qquad \square$$

C Guaranteed information at positive loss

This section uses the full class of finite communicating δ-admissible models, for a fixed $0 \leq \delta < 1/2$, with common visible labels and possible initial observation. There is no common upper bound on the number of hidden states. Observations may be finite or countable and discrete. We first construct a finite intermediate realization, then explain why equal laws need not give a common oracle directly.

C.1 Conditioning and finite common realizations

With the conditional values $V^{\pi}_{M,s,o}(g)$ and $V^{*}_{M,s,o}(g)$ of Section 3, a policy is conditionally competent if it meets the relative guarantee for every such conditional context. This property will be obtained by selecting an oracle, rather than imposed on all supplied oracles. Prefixing a goal with a current observation gate gives the native goal $g[o]$.

Lemma C.1 (Normalization by the initial observation)**.** *If $\pi \in \mathcal{C}_{\delta}(M)$, then the complete policy*

$$(\mathsf{N}\pi)(h, g) = \pi(h, g[o_0])$$

is conditionally δ-competent and belongs to $\mathcal{C}_{\delta}(M)$.

Proof. We use the initial gate to isolate one conditional value. Its success is zero at every other initial observation, so competence for $g[o]$ gives

$$\Omega(o \mid s) V^{\pi_{g[o]}}_{M,s,o}(g) \geq (1-\delta)\Omega(o \mid s) V^{*}_{M,s,o}(g).$$

The initial observation is retained in every history, so $\mathsf{N}\pi$ follows this gated policy throughout the trajectory. Cancelling the positive factor gives conditional competence. Averaging over the initial observation and bounding a supremum of a sum by the sum of suprema gives

$$V^{\mathsf{N}\pi}_{M,s}(g) \geq (1-\delta)\sum_{o} \Omega(o \mid s) V^{*}_{M,s,o}(g) \geq (1-\delta) V^{*}_{M,s}(g). \qquad \square$$

We use the controlled test functions, predictive polytopes and prefix operators defined in Section 3. Test evaluations separate vectors in their finite-dimensional span, so the equality constraints below require only finitely many scalar linear equations.

The following proof of Lemma 3.7 constructs the realization.

Proof of Lemma 3.7. We separate a common predictive function into an emission law and finitely many action-dependent successor mixtures. By the Minkowski–Weyl theorem, every nonempty

bounded polyhedron $\{z \geq 0 : Bz = b\}$ is finitely generated (Ziegler, 1995, Section 1.1). Applying this fact to the two simplex representations of an intersection shows that

$$\mathcal{P}_M \cap \mathcal{P}_N = \operatorname{conv}\{x_1, \ldots, x_k\}.$$

We choose decompositions $x_j = \sum_s u_{j,s} f_{M,s} = \sum_t v_{j,t} f_{N,t}$ and put $p_j(o) = x_j(o)$. When $p_j(o) > 0$, the normalized successor $\mathcal{L}_{o,a} x_j / p_j(o)$ belongs to the intersection for every action. Its endpoint posterior weights are

$$d^M_{j,o,s} = u_{j,s} \Omega_M(o \mid s)/p_j(o), \qquad d^N_{j,o,t} = v_{j,t} \Omega_N(o \mid t)/p_j(o).$$

We use the bounded polytope of tuples $(d^M, d^N, (\psi_a)_a)$ satisfying simplex constraints and

$$\sum_s d^M_s q_{M,s,a} = \sum_{i=1}^{k} \psi_{a,i} x_i = \sum_t d^N_t q_{N,t,a} \qquad (a \in \mathcal{A}). \tag{15}$$

It contains a tuple for each (j, o) with $p_j(o) > 0$. We choose finite generators indexed by $\ell = 1, \ldots, l$ and decompose these tuples with weights $\lambda_{j,o,\ell}$. We define

$$w_{j,\ell}(o) = p_j(o) \lambda_{j,o,\ell}, \qquad b_{j,\ell} = \sum_o w_{j,\ell}(o),$$

using zero weights when $p_j(o) = 0$. The model has the finitely many states (j, ℓ) with $b_{j,\ell} > 0$, emissions and transitions

$$\Omega_R(o \mid j, \ell) = \frac{w_{j,\ell}(o)}{b_{j,\ell}}, \qquad P_R((i, r) \mid (j, \ell), a) = \psi^\ell_{a,i} b_{i,r}. \tag{16}$$

Both are normalized. The transition is independent of the observation emitted at its source, as required for a state-emitting model. Let b_j place mass $b_{j,\ell}$ on (j, ℓ). Induction on test length, using (15), gives

$$f_{R,b_j}(oaw) = \sum_\ell w_{j,\ell}(o) \sum_i \psi^\ell_{a,i} f_{R,b_i}(w) = p_j(o) \sum_\ell \lambda_{j,o,\ell} \sum_i \psi^\ell_{a,i} x_i(w) = x_j(oaw).$$

The single-observation and empty tests initialize the induction. Mixtures of these initial distributions realize the intersection.

We now transfer competence. At a state (j, ℓ) and observation with $w_{j,\ell}(o) > 0$, positivity in the convex decomposition implies that $d^{M,\ell}_s > 0$ only if $\Omega_M(o \mid s) > 0$. Conditional on this observation, the controlled future in R is the mixture with weights $d^{M,\ell}$ of the endpoint conditional futures. Equality for finite tests extends to every fixed feedback policy by its common action factors, and to native success by increasing finite acceptance events. Thus

$$V^\sigma_{R,(j,\ell),o}(g) = \sum_s d^{M,\ell}_s V^\sigma_{M,s,o}(g), \qquad V^*_{R,(j,\ell),o}(g) \leq \sum_s d^{M,\ell}_s V^*_{M,s,o}(g).$$

A conditionally competent endpoint policy attains the required fraction of the latter upper bound. Averaging over the initial emission transfers statewise competence. For either endpoint $i \in \{M, N\}$,

$$V^\sigma_{R,(j,\ell)}(g) \geq (1 - \delta) \sum_o \Omega_R(o \mid j, \ell) V^*_{R,(j,\ell),o}(g) \geq (1 - \delta) V^*_{R,(j,\ell)}(g). \qquad \square$$

We now impose communication while preserving the common law.

Lemma C.2 (A communicating restriction with the common law)**.** *If $J_M = J_N$, the realization in Lemma 3.7 has a closed communicating restriction R_* with that same random-walk law. The competence transfers remain valid there.*

Proof. We use stationary empirical frequencies to rule out components with different random-walk laws. Equal J gives a common stationary controlled test function

$$\phi = \sum_s \eta_M(s) f_{M,s} = \sum_t \eta_N(t) f_{N,t} \in \mathcal{P}_M \cap \mathcal{P}_N.$$

We choose an initial distribution b in R realizing ϕ. Under uniform actions it generates the random-walk law J. If Q is the averaged hidden transition matrix of R, a limit point ξ of the Cesàro averages

of bQ^t satisfies $\xi Q = \xi$ by telescoping. Each time shift preserves J, so continuity and linearity of finite cylinder probabilities show that ξ also generates J.

A stationary distribution of a finite chain is a mixture of the stationary distributions on its closed communicating classes. We write the corresponding output decomposition as $J = \sum_C \omega_C J_C$, retaining positive weights. This follows from mass balance on the finite acyclic graph of communicating classes: a nonclosed class with stationary mass would give positive outward flow after its predecessors are removed. Each retained class is closed.

For every fixed cylinder, Lemma B.1 in M gives empirical frequency convergence in probability to $J(c)$ under J. The decomposition implies the same convergence under each J_C, since its error probability is at most the mixture error divided by ω_C. Frequencies are bounded by one, so their expectations converge too. Under stationary J_C those expectations equal $J_C(c)$ for every length. Thus $J_C(c) = J(c)$ for all cylinders. We choose any retained class. Closedness under the uniform action average implies closedness under every action, so the original controlled futures and competence inequalities at its states are preserved. Its random-walk law satisfies

$$J_{R_*}(c) = J_C(c) = J(c) \qquad (c \text{ any cylinder}). \qquad \square$$

Proof of Theorem 3.6. We first suppose $J_M = J_N$ and choose competent endpoint policies. Lemma C.1 makes each conditionally competent. Lemmas 3.7 and C.2 provide one finite communicating R to which both transfer. Each policy is shared by its original model and R, making R admissible. The common observation is possible because its marginal in J_R is positive. Conversely, a common competent oracle on each pair forces $J_M = J_R = J_N$ by Lemma 3.3, or its countable version.

If a decoder D works for every competent oracle and target F, choose the two shared policies σ_M, σ_N. Then $F(M) = D(\sigma_M) = F(R) = D(\sigma_N) = F(N)$, so F is constant on a fiber of J. Conversely each admitted oracle determines a unique J by separation. A function $\tilde{F}$ of this law defines a decoder with

$$D(\pi) = \tilde{F}(J_M) = F(M) \qquad (\pi \in \mathcal{C}_\delta(M)). \qquad \square$$

The root-distribution function has the same guaranteed information: Corollary D.2 shows it determines J, and every target determined by it is also determined by the full policy. The theorem uses bridges of unrestricted size. At positive loss, it does not classify the guaranteed information in every subclass with a fixed number of states.

C.2 Equal laws can have disjoint competent-oracle sets

Fix $0 < \alpha_0 < 1/2$ and let $p_- = \alpha_0$, $p_+ = 1 - \alpha_0$. Hidden states are $(i, t) \in \{0, 1\} \times \{-, +\}$ and emit i. For actions L, R, model M_{α_0} has

$$P_M((j, v) \mid (i, t), \mathsf{L}) = \tfrac{1}{2} p_t^j (1 - p_t)^{1-j}, \qquad P_M((j, v) \mid (i, t), \mathsf{R}) = \tfrac{1}{4}. \tag{17}$$

Model N_{α_0} exchanges the two actions. A reference model R_0 has two observed states and transitions to either state with probability $1/2$ under either action. All these transitions are positive. Under a uniform initial hidden law, each action refreshes a uniform type independently of a fair observation. All three models have the same random-walk law, with fair IID observations and independent uniform actions.

Lemma C.3 (Two-context relative loss)**.** *One action has values l, u in two contexts, where $0 < l < u$, and the other has a common value $q \geq 0$. The minimum largest relative loss is zero for $q \notin (l, u)$, while for $l < q < u$ it equals*

$$\mathcal{R}(l, u; q) = \frac{(1 - l/q)(1 - q/u)}{2 - l/q - q/u}, \qquad \sup_{l<q<u} \mathcal{R}(l, u; q) = \frac{1 - \sqrt{l/u}}{2}. \tag{18}$$

The supremum is attained at $q = \sqrt{lu}$. Coincident values and zero benchmarks have a common maximizing action.

Proof. We choose a mixing probability to balance the two losses. Outside (l, u) one action maximizes both contexts. Inside, put $a = 1 - l/q$ and $b = 1 - q/u$; using the first action with probability θ

gives losses θa and $(1-\theta)b$. Their maximum is minimized at $\theta = b/(a+b)$, with value $ab/(a+b)$. For $c = l/u$ and $z = l/q + q/u$, the arithmetic-geometric mean inequality gives $z \geq 2\sqrt{c}$ and $z < 2$. Consequently

$$\frac{1-\sqrt{c}}{2} - \frac{1-z+c}{2-z} = \frac{(1+\sqrt{c})(z-2\sqrt{c})}{2(2-z)} \geq 0, \qquad z = 2\sqrt{c} \text{ at } q = \sqrt{lu}. \qquad \square$$

Proof of Proposition 3.5. We select the four-state models at their exact admissibility radius. After the first transition, the type is a fresh uniform bit, independent of the next observation. The future controlled process is therefore the fair reference process. For any goal and fixed initial observation, Lemma B.2 supplies optimal continuations after either action. If their values following next observations zero and one on the L branch are z_0, z_1, its hidden-type values are $u_t = (1-p_t)z_0 + p_t z_1$. The R branch has one common value q.

Set $l = \min(u_-, u_+)$ and $u = \max(u_-, u_+)$. If the nonzero values differ, then

$$l \geq \alpha_0(z_0 + z_1), \qquad u \leq (1-\alpha_0)(z_0+z_1), \qquad l/u \geq \alpha_0/(1-\alpha_0).$$

Lemma C.3 gives a root mixture attaining largest loss at most $r_{\alpha_0} = (1 - \sqrt{\alpha_0/(1-\alpha_0)})/2$. The mixture depends on the initial observation and goal, followed by the common optimal continuations. Degenerate branch values have a common maximizing action. This defines one complete statewise policy at that loss.

The radius is attained from below by finite goals. Gate on an initial observation and compare the L branch $O_1 = 1$ with an R branch accepting any selected k of the 2^d fair observation words, with all continuation actions prescribed as R. Their values are $(\alpha_0, 1-\alpha_0)$ and the common dyadic probability $k/2^d$. Dyadic probabilities approach $\sqrt{\alpha_0(1-\alpha_0)}$, and (18) is continuous there. No smaller loss works for all these goals. The action-exchanged model has the same radius.

Every competent oracle of either four-state model is competent in R_0: averaging its two initial-type inequalities gives at least $(1-\delta)$ times the average of their optima, which bounds the reference optimum from above. We now choose $\alpha_0 = \Delta^2/(1+\Delta^2)$, with the margin Δ of Section 4, so $r_{\alpha_0} = \delta$. Each four-state model shares its oracle with R_0.

For the goal requiring $O_0 = 0$ and then $O_1 = 1$, let θ be the probability of L. At initial hidden type $-$, competence in M_{α_0} requires $\theta \leq \delta/(1-2\alpha_0)$, while in N_{α_0} it requires $\theta \geq 1 - \delta/(1-2\alpha_0)$. These conflict because

$$\frac{1-2\alpha_0}{2} - \delta = \frac{\Delta(1-\Delta)^2}{2(1+\Delta^2)} > 0. \qquad \square$$

D Recovery from actions

The margin Δ, the error level e_0, the query family $g(c, N, q, w)$ and the cylinder length d are those of Section 4; $\gamma > 0$ is a comparison tolerance.

D.1 Proof of Lemma 4.2

The proof below uses competence at every hidden start, or under one fixed initial law; Proposition D.4 extends it to the average condition (9).

Proof. We average the controlled event probabilities over the random word. Conditional on $O_0 = o_*$, let A_w be the optimal branch value after a_+ and B_w that after a_-. These are the probabilities of E_w and E_w^c from the respective post-marker hidden laws, since every surviving continuation must follow w. If x_w, y_w are the two root probabilities, cancellation of the initial-observation gate in the competence inequality gives

$$x_w \geq (1-\delta)A_w - B_w, \qquad y_w \geq (1-\delta)B_w - A_w. \tag{19}$$

For statewise competence we may take any positive hidden mixture before conditioning: its policy value is at least $(1-\delta)$ times the mixture optimum. This also shows that the argument applies to competence under one fixed initial law with positive mass on o_*.

Uniform W gives independent uniform measured actions in expectation. Lemma B.1 and (10) bound the error probability by e_0 from either post-marker law. Above the threshold, $\mathbb{E}A_W \geq 1 - e_0$ and $\mathbb{E}B_W \leq e_0$, so $\mathbb{E}x_W \geq 1/2 + \Delta/4$. Below it, the same argument gives $\mathbb{E}y_W \geq 1/2 + \Delta/4$, and $x_W + y_W \leq 1$ gives the other bound. The independent words and fresh action draws form bounded independent tests. Their exponential moment bound yields

$$\mathbb{P}(\text{incorrect majority} \mid \text{earlier queries}) \leq \exp\bigl(-2K(\Delta/4)^2\bigr) = \exp(-K\Delta^2/8). \qquad \square$$

D.2 Accuracy and pointwise convergence

Theorem D.1 (Pointwise recovery without a mixing bound)**.** *Enumerate the nonempty pair cylinders as $c_1, c_2, \ldots$. At stage $t \geq 2$, estimate the first t cylinders by bisection with*

$$N_t = t^4, \qquad B_t = \lceil \log_2 t \rceil, \qquad K_t = \left\lceil \frac{8}{\Delta^2} \log(2t^3 B_t) \right\rceil. \tag{20}$$

For every model and competent oracle, almost surely each fixed cylinder eventually satisfies $|\widehat{J}_t(c_j) - J_M(c_j)| \leq 3/(2t)$. With zero estimates on unqueried coordinates,

$$\widehat{J}_t(c_j) \longrightarrow J_M(c_j) \quad \forall j, \qquad \sum_{j \geq 1} 2^{-j} |\widehat{J}_t(c_j) - J_M(c_j)| \longrightarrow 0. \tag{21}$$

Each stage uses at most tB_tK_t root replies. No model parameters enter the schedule.

Proof of Theorem 4.3. We set $\gamma = \epsilon/2$ so that comparison error and interval width share the accuracy budget. Lemma 4.2 and a union bound make all comparisons at distance at least γ from the truth correct with probability at least $1 - \zeta$. On this event, the maintained lower endpoint never exceeds $J_M(c) + \gamma$ and the upper endpoint never falls below $J_M(c) - \gamma$. Comparisons inside the band may be arbitrary. The final interval has width 2^{-B}, so its midpoint satisfies

$$|\widehat{J}(c) - J_M(c)| \leq \gamma + 2^{-B-1} \leq \epsilon. \qquad \square$$

Proof of Theorem D.1. We choose the length scale to dominate every fixed Poisson constant. For cylinder c_j and band $1/t$, (12) bounds the error by $16G^2_{M,c_j}/t^2$, which is eventually at most e_0. Among all cylinders for which this inequality holds, the chance of any erroneous comparison outside the band at stage t is at most

$$tB_t \exp(-K_t\Delta^2/8) \leq \frac{1}{2t^2}.$$

The tail sums tend to zero, so almost surely only finitely many such stage errors occur. Each fixed cylinder eventually meets the bound; the endpoint argument in Theorem 4.3 gives error at most $t^{-1} + 2^{-B_t - 1} \leq 3/(2t)$. For a finite prefix of k coordinates the errors vanish together. The remaining weighted error is at most 2^{-k}, giving

$$\limsup_{t \to \infty} \sum_{j \geq 1} 2^{-j} |\widehat{J}_t(c_j) - J_M(c_j)| \leq 2^{-k} \longrightarrow 0. \qquad \square$$

D.3 Variants of the interface

Corollary D.2 (Extensions of the root-action construction)**.** *Theorems 4.3 and D.1 hold under each of the following variants.*

1. *Competence is evaluated under one fixed model-specific initial law b, with $\mathbb{P}_{M,b}(O_0 = o_*) > 0$, instead of separately at all hidden starts.*
2. *The observations are a known effectively enumerable countable discrete alphabet. Each queried cylinder is represented by its finitely many named labels and their complement.*
3. *Queries return the exact root distribution; the a_+ probability is used as a bounded sample. A fixed selected optimal action is sufficient when $\delta = 0$.*

4. *At stage t a different policy is supplied, provided it meets the same loss guarantee for that stage's goals and returns fresh conditional action draws. All those goals can have common exact depth $t^4 + \max_{j \le t} |c_j|$.*

Proof. We identify the inputs used by the threshold bound. For the first variant, the initial gate makes the value zero outside $O_0 = o_*$. Conditional on that observation, the branch values in (19) of Appendix D.1 use the posterior under b. Cancelling its positive probability gives the same inequalities. Their averaged bounds are uniform over both post-marker hidden distributions.

For countable observations, the finite partition used in Appendix B.2 still yields a finite irreducible window chain. Each goal has finitely many cell words, and effective enumeration gives a cylinder schedule. No infinite disjunction is needed.

For exact distributions, the returned coordinate lies in $[0, 1]$ with mean $\mathbb{E}_W x_W$. We choose fresh independent words so that concentration applies to these bounded variables. A selected exact action is their indicator-valued special case.

For policies that vary by stage, we condition on all previous queries. The threshold bound is uniform over the competent policy used at the current stage. We pad shorter disjuncts with always-true current tests, since these consume no additional time and leave the event unchanged. The summable bound used in Theorem D.1 is still

$$\mathbb{P}(\text{an erroneous qualified comparison at stage } t) \le tB_t \exp(-K_t \Delta^2/8) \le \frac{1}{2t^2}. \qquad \square$$

For the first variant, the identifiability direction survives, but equality of random-walk laws need not imply equality of oracle sets. The converse in Theorem 3.1 uses exact statewise admissibility and positive stationary weights. For a noneffectively specified countable alphabet, the complete-function information theorem below remains meaningful, while the sampling algorithm requires an enumeration interface.

Corollary D.3 (Countable exact information)**.** *Theorem 3.1 and Corollary 3.4 hold for finite or countable discrete observations, with arbitrary subset predicates in the native goal language and a common possible initial observation.*

Proof. We retain the same positive-mixture identity, since its sum is over the finite hidden-state set. At every finite time the transcript set is at most countable. Equality in (4) therefore gives equality of finite-time acceptance probabilities by countable additivity, and equality of full goal values by continuity from below. The finite partition argument in Corollary D.2 preserves the separation direction. Applying the decoder argument of Corollary 3.4 yields

$$J_M = J_N \iff \mathcal{C}_0(M) = \mathcal{C}_0(N) \iff \mathcal{C}_0(M) \cap \mathcal{C}_0(N) \neq \varnothing. \qquad \square$$

D.4 Competence on the query distribution and finite stopping

D.4.1 Average competence on prescribed words

For the fixed initial law b in (9), put $\rho = \mathbb{P}_{M,b}(O_0 = o_*) > 0$. The queried goal g_W uses a fresh uniform prescribed word W conditional on preceding calls $\mathcal{F}$.

Proposition D.4 (Recovery under average competence on the queried goals)**.** *Lemma 4.2, Theorem 4.3, and Theorem D.1 hold under* (9) *for their queried threshold families. The policy may change between calls if the conditional bound continues to hold.*

Proof. We average the competence inequality before extracting the root preference. Conditional on the gate, let A_W and B_W be the success probabilities of the two prescribed branches and let p_W be the root probability of the positive marker. The optimal conditional value is $\max(A_W, B_W)$; every continuation of the supplied policy has value at most $p_W A_W + (1 - p_W)B_W$. Suppose $J_M(c) \ge q + \gamma$, and put $e = 16G^2_{M,c}/(N\gamma^2)$. Uniform averaging over both post-marker hidden laws gives $\mathbb{E}A_W \ge 1 - e$ and $\mathbb{E}B_W \le e$. We cancel the gate probability in (9) to obtain

$$\mathbb{E}p_W + e \ge \mathbb{E}[V_b^\pi(g_W)/\rho] \ge (1 - \delta)\mathbb{E}\max(A_W, B_W) \ge (1 - \delta)(1 - e).$$

All expectations may be conditional on $\mathcal{F}$. The length in Lemma 4.2 makes $e \le e_0$, and hence $\mathbb{E}p_W \ge 1/2 + \Delta/4$. Exchanging the two branches gives $\mathbb{E}p_W \le 1/2 - \Delta/4$ when $J_M(c) \le q - \gamma$.

For changing policies, we use the conditional margin and the bounded martingale-difference form of Hoeffding's inequality. The comparison error is at most $\exp(-K\Delta^2/8)$. The same bisection and summable error arguments therefore give

$$\mathbb{P}\big(|\widehat{J}(c) - J_M(c)| > \epsilon\big) \le \zeta, \qquad \widehat{J}_t(c) \longrightarrow J_M(c) \quad \text{almost surely.} \qquad \square$$

D.4.2 Reusing a finite menu of prescribed words

For a fixed cylinder and length, the count law for one prescribed word answers every integer cutoff. This permits reuse across an adaptive search.

Proposition D.5 (One reply per word and cutoff)**.** *Fix a policy and a public menu $\mathcal{M}$ of threshold families with common (c, N), each satisfying the unconditional form of* (9)*. Assume the length bound of Lemma 4.2 for tolerance γ. Draw K independent uniform prescribed words once. For each queried menu entry, obtain one fresh root action per word. If*

$$K \ge \left\lceil 8\Delta^{-2} \log(|\mathcal{M}|/\zeta) \right\rceil, \tag{22}$$

then, with probability at least $1 - \zeta$, every adaptively queried majority whose cutoff is more than γ from $J_M(c)$ has the correct sign. For B comparisons there are at most BK replies, but only K distinct count-law computations for the cylinder. Integer cutoffs $q = k/N$, $0 \le k \le N$, give $|\mathcal{M}| = N + 1$.

Proof. We draw the entire array of possible replies before deciding which columns to reveal. In column k, reply X_{ik} uses word W_i and independent private action randomness. For each fixed k, these replies are independent across i, with the unconditional margin proved above. Different columns may be dependent because they share words. Hoeffding's inequality bounds each qualified column's error by $\exp(-K\Delta^2/8)$. On the event that every qualified column is correct, adaptive selection can reveal only correct qualified answers. Taking a union bound over the public menu gives

$$\mathbb{P}(\text{any erroneous qualified answer}) \le |\mathcal{M}| \exp(-K\Delta^2/8) \le \zeta. \qquad \square$$

The computation count concerns evaluation of a fixed synthetic planner: its complete count distribution can be cached once per word, then queried by tail sums. The original fresh-word schedule uses only its visited cutoffs in the confidence bound. Reuse instead pays for the full menu inside the logarithm in (22).

D.4.3 Why an unknown mixing rate prevents a finite certificate

Theorem D.6 (No uniformly valid finite stopping rule)**.** *Fix $0 < \delta < 1/2$, $0 < \epsilon < 1/4$, and $0 < \zeta < 1/2$. Consider root access for finite current/next goals, with competence under one fixed initial law. On the communicating class with at most two hidden states, no algorithm can both stop almost surely and estimate $\mathbb{P}_{\eta_M}(O_0 = 1)$ within ϵ with probability at least $1 - \zeta$, uniformly over the models and their δ-competent policies. This holds even when each call returns the exact root action distribution.*

Proof. We hide a different random-walk law behind an arbitrarily rare transition. The reference M_0 has one state, observation probability $p_0 = 1/4$, and identical actions. We choose a deterministic exact policy π_0 for each finite goal by backward induction on its finite history tree. For $0 < \lambda < 1$, the alternative starts in state zero, has observation-one probabilities $(1/4, 3/4)$, and uses the same kernel for every action:

$$P_\lambda = \begin{pmatrix} 1-\lambda & \lambda \\ \lambda^2 & 1-\lambda^2 \end{pmatrix}, \qquad \eta_\lambda = \frac{(\lambda, 1)}{1+\lambda}, \qquad p_\lambda - p_0 = \frac{1}{2(1+\lambda)}. \tag{23}$$

Both models admit the root gate $o_* = 0$.

We couple their observations and policy randomness until the alternative first leaves state zero. For any policy and a goal decided within H transitions, the success probabilities differ by at most $H\lambda$.

Every satisfiable such goal has reference optimum at least $b_H = 4^{-(H+1)}$: one satisfying observation–action word suffices, and each observation has probability at least $1/4$. Unsatisfiable goals have zero value in both models. Thus, when $\lambda \le \delta b_H/[(2-\delta)H]$, the reference policy satisfies

$$V_\lambda^{\pi_0}(g) - (1-\delta)V_\lambda^*(g) \ge \delta V_0^*(g) - (2-\delta)H\lambda \ge \delta b_H - (2-\delta)H\lambda \ge 0.$$

We define a complete policy π_λ equal to π_0 on every goal of horizon at most H, and optimal under the alternative's initial law on every longer finite goal. It is δ-competent throughout the finite goal language. Its complete root distributions agree with the reference for every goal of horizon at most H.

Suppose a uniformly valid, almost surely terminating algorithm exists. Under (M_0, π_0) its largest queried horizon is finite almost surely. We choose $0 < \xi < 1 - 2\zeta$ and then $H \ge 1$ so that all its queries have horizon at most H with probability at least $1 - \xi$. With probability at least $1 - \zeta - \xi$, it also returns an estimate within ϵ of p_0. We now choose λ small enough for the competence bound above and $p_\lambda - p_0 > 2\epsilon$. Coupling the algorithm's randomness makes its entire transcript and answer identical on that event in the two models. The answer is then wrong for the alternative, so

$$\mathbb{P}_{M_\lambda, \pi_\lambda}(|\widehat{p} - p_\lambda| > \epsilon) \ge 1 - \zeta - \xi > \zeta. \qquad \square$$

The premise is competence under the fixed initial law of Theorem 4.1 and Corollary D.2. Statewise competence would also constrain the alternative from state one, where the optimal first action can differ from the reference policy's, and the theorem does not settle whether a finite certificate exists under that premise.

A supplied upper bound on the Poisson constant restores a finite guarantee through Theorem 4.3 and Proposition H.2. Without that promise, Theorem D.1 supplies an almost surely convergent sequence of estimates.

E Competence and fixed physical horizons

E.1 Loss one half

Proof of Proposition 4.4(i). We mix complete exact policies with one persistent random choice. Model M has observed states x, y; action a toggles and action b stays. Model N exchanges these action effects. Each model admits an exact statewise policy for every native goal, by full observation and Lemma B.2. We choose such complete policies π_M, π_N.

For a history h_t and goal g, define $w_i(h_t, g) = \prod_{j<t} \pi_i(a_j \mid h_j, g)$ for $i \in \{M, N\}$. When the denominator is positive, set

$$\pi(a \mid h_t, g) = \frac{w_M(h_t, g)\pi_M(a \mid h_t, g) + w_N(h_t, g)\pi_N(a \mid h_t, g)}{w_M(h_t, g) + w_N(h_t, g)}. \tag{24}$$

At a zero denominator we choose an arbitrary distribution. These outputs depend only on the common history. The action-probability products telescope to $(w_M + w_N)/2$. Multiplication by the environment's controlled observation likelihood therefore gives, in either fixed environment, the equal mixture of the trajectory laws under π_M and π_N. Each policy receives weight one half for the whole trajectory. Thus

$$V_{i,s}^{\pi}(g) = \tfrac{1}{2}V_{i,s}^{\pi_M}(g) + \tfrac{1}{2}V_{i,s}^{\pi_N}(g) \ge \tfrac{1}{2}V_{i,s}^*(g) \quad (i \in \{M, N\}).$$

Uniform actions make the hidden stationary distribution uniform in both models, but their observable cylinders differ:

$$J_M(O_0 = \mathsf{x}, A_0 = \mathsf{a}, O_1 = \mathsf{y}) = \tfrac{1}{4}, \qquad J_N(O_0 = \mathsf{x}, A_0 = \mathsf{a}, O_1 = \mathsf{y}) = 0. \qquad \square$$

E.2 A common policy for every bounded-horizon goal

For $0 < \lambda, p < 1$, consider two fully observed states with the same kernel under every action:

$$P_{\lambda,p} = \begin{pmatrix} 1 - \lambda p & \lambda p \\ \lambda(1-p) & 1 - \lambda(1-p) \end{pmatrix}. \tag{25}$$

The stationary mass of state one is p. We allow every Boolean event of a trajectory with at most H transitions, including all its actions. This contains the native goals whose outcomes are determined within that horizon.

Lemma E.1 (A parameter-independent bounded-horizon policy)**.** *For $0 < r_0 < 1/2$, there is one complete deterministic policy π^{lex}, independent of λ, p, that satisfies every such goal at every residual history with value ratio at least*

$$\frac{(1-\lambda)^H}{1+2^H\lambda/r_0^H} \quad \textit{for all } p \in [r_0, 1-r_0]. \tag{26}$$

In particular, $\lambda = \delta/(H + 2^H/r_0^H)$ makes this policy δ-competent simultaneously throughout that interval.

Proof. We rank successful future state words by their numbers of switches. At a residual history ending in state i with $L \le H$ transitions left, any deterministic policy has a count vector $(c_0, \dots, c_L)$, where c_k counts its successful words with k switches. We choose a policy maximizing this vector lexicographically, with smaller switch counts considered first.

A common choice can be made consistently at every residual history by backward induction. Represent a subtree by the formal polynomial $\sum_k c_k z^k$. For a selected action at state i, the two successor subtrees combine as $C_i(z) + zC_{1-i}(z)$. Addition and multiplication by z preserve lexicographic nonnegativity. Thus choosing a maximizing continuation in each subtree and then a maximizing action gives a consistent maximum, with a fixed action order at ties. The count polynomials depend on the goal, but not on the transition parameters. After the goal is decided we prescribe the first action in that fixed order.

For a word with k switches, its probability is $\lambda^k w_{i,k}(p)D$, where

$$w_{0,k}(p) = p^{\lceil k/2 \rceil}(1-p)^{\lfloor k/2 \rfloor}, \qquad w_{1,k}(p) = p^{\lfloor k/2 \rfloor}(1-p)^{\lceil k/2 \rceil}, \qquad (1-\lambda)^H \le D \le 1.$$

The directions of successive binary switches alternate, which gives their common leading weight. On the specified parameter interval, $w_{i,k}(p) \ge r_0^H$.

If the optimum is zero there is no value to lose. Otherwise we choose the first achievable successful switch count k and its maximal count $c \ge 1$. The lexicographic policy attains at least $c\lambda^k w_{i,k}(p)(1-\lambda)^H$. A competitor has no successes at fewer switches, at most c at k switches, and at most 2^H remaining words, each with probability at most λ^{k+1}. Randomized policies are mixtures of deterministic finite trees. The ratio is consequently at least (26). With $\kappa = 2^H/r_0^H$ and the specified λ, Bernoulli's inequality gives

$$\frac{(1-\lambda)^H}{1+\kappa\lambda} \ge \frac{1-H\lambda}{1+\kappa\lambda} = 1 - \frac{(H+\kappa)\lambda}{1+\kappa\lambda} = 1 - \frac{\delta}{1+\kappa\lambda} > 1-\delta. \qquad \square$$

Proof of Proposition 4.4(ii). We keep one complete bounded-horizon policy while moving the stationary target towards opposite endpoints. For $r_0 \in (0, 1/2)$, take the common λ from Lemma E.1 and the models P_{λ,r_0} and $P_{\lambda,1-r_0}$. Both communicate and have the same permitted policy function π^{lex}.

We extend each oracle separately to the remaining native goals. From each observed initial state and each such goal, a positive supremum admits a policy attaining more than its $(1-\delta)$ fraction, since $\delta > 0$; at a zero supremum any policy suffices. The initial state is observed, so these choices combine into one statewise competent goal-conditioned function. They leave the common bounded-horizon restriction unchanged.

An arbitrary identifier, including adaptive or randomized ones, thus has one output distribution $\widehat{p}$ in the two models. The triangle inequality gives

$$\max\{\mathbb{E}|\widehat{p} - r_0|, \mathbb{E}|\widehat{p} - (1-r_0)|\} \ge \frac{1-2r_0}{2}.$$

We let r_0 tend to zero, retaining positive transition probabilities at every member. The constant estimate $1/2$ supplies the upper bound throughout the class, so

$$\frac{1}{2} \le \inf_{\widehat{p}} \sup_{M,\pi} \mathbb{E}|\widehat{p} - p_M| \le \frac{1}{2}. \qquad \square$$

The result excludes a uniform bound at fixed physical horizon. It allows pointwise recovery with growing horizons and does not assert an exact-oracle lower bound at $\delta = 0$. The initial state labels supplied in a query are inputs, rather than samples of the stationary hidden distribution.

F PREDICTIVE REALIZATION AND HIDDEN PARAMETERS

The recovered random-walk law has the classical finite-rank representation used in PSRs and observable operator models (Littman et al., 2001; Jaeger, 2000; Thon & Jaeger, 2015). Define the Hankel matrix and its one-letter insertions by $\mathsf{H}(u,v) = J(uv)$ and $\mathsf{H}_x(u,v) = J(uxv)$ for pair words u, v. Equation (2) factors H through the hidden-state space, so its rank r is at most $|S|$.

Corollary F.1 (The recovered transition system)**.** *Choose sets U, V of r words such that $H_0 = \mathsf{H}(U,V)$ is invertible. The law recovered by Theorem D.1 determines*

$$\alpha = \mathsf{H}(\varnothing, V), \qquad \beta = H_0^{-1}\mathsf{H}(U,\varnothing), \qquad T_x = H_0^{-1}\mathsf{H}_x(U,V), \tag{27}$$

with $J(x_1 \cdots x_k) = \alpha T_{x_1} \cdots T_{x_k}\beta$. This is a minimal linear representation, unique up to one invertible change of coordinates. If a state cap n is given, words of length at most $2n-1$ suffice for its construction; sampled estimates consistently recover the rank and one fixed realization.

Prediction from an observable history. For a pair history h with $J(h) > 0$, the prediction row is $v_h = \mathsf{H}(h,V)/J(h)$. Its transition and output equations are

$$v_{hx} = \frac{v_h T_x}{v_h T_x \beta}, \qquad \mathbb{P}_J(x \mid h) = v_h T_x \beta, \tag{28}$$

when the denominator is positive. Removing the independent action factors gives controlled suffix probabilities: for $x_j = (o_j, a_j)$ they equal $m^k v_h T_{x_1} \cdots T_{x_k}\beta$. These predictions use the stationary initialization defining the random-walk law. The operators need not have nonnegative entries.

Lemma F.2 (Finite spans and equality of random-walk laws)**.** *For a model with at most n hidden states, prefix and suffix words of length at most $n-1$ attain the full Hankel rank $r \le n$. For models with at most n_M and n_N states,*

$$J_M = J_N \quad \Longleftrightarrow \quad J_M(w) = J_N(w) \quad \textit{for every } |w| \le n_M + n_N - 1. \tag{29}$$

The bounded-length family is finite when the observation alphabet is finite.

Proof. We expand the reachable row span and the observable column span until each is invariant under every letter matrix. Let $\mathcal{R}_k$ be the span of $\eta_M B_u$ for $|u| \le k$, and let $\mathcal{V}_k$ be the span of $B_v\mathbf{1}$ for $|v| \le k$. Both begin with dimension one and have dimension at most n. If $\mathcal{R}_k = \mathcal{R}_{k+1}$, multiplication by any B_x stays in $\mathcal{R}_k$, so later expansion adds nothing. The column argument is the same with left multiplication. Each strict expansion adds at least one dimension, so both spans stabilize by $k = n-1$.

The factorization $\mathsf{H}(u,v) = (\eta_M B_u)(B_v\mathbf{1})$ gives the rank bound. Since all factor rows and columns are spanned at length $n-1$, that truncation attains the full Hankel rank. Finite bases can also be selected when the alphabet is countable.

For equality, we use block-diagonal letter matrices $\mathrm{diag}(B_x^M, B_x^N)$, initial row $(\eta_M, -\eta_N)$, and terminal column $(\mathbf{1}, \mathbf{1})^\top$. Their word weight is $J_M(w) - J_N(w)$. The column span stabilizes by length $n_M + n_N - 1$, so a row annihilating all columns through that length annihilates every word column:

$$J_M(w) - J_N(w) = 0 \quad (|w| \le n_M + n_N - 1) \implies J_M(w) - J_N(w) = 0 \quad (w \in \Sigma^*). \qquad \square$$

This is the finite-span argument underlying classical automaton and observable-system equivalence. The broader common algebra, rank minimality, and similarity relation are developed by Thon & Jaeger (2015, Section 2).

Algebraic part of Corollary F.1. We express every Hankel row in the basis indexed by U. Since its restriction H_0 to V is invertible and the full rank is r, the coefficient row for prefix h is $\mathsf{H}(h,V)H_0^{-1}$. Evaluating at suffixes xv, with $v \in V$, gives

$$\mathsf{H}(hx, V) = \mathsf{H}(h,V)H_0^{-1}\mathsf{H}_x(U,V) = \mathsf{H}(h,V)T_x. \tag{30}$$

Evaluation at the empty suffix gives $J(h) = \mathsf{H}(h,V)\beta$. Iterating (30) from the empty prefix yields the formula in Corollary F.1. Any linear representation of dimension k factors the Hankel matrix through a k-dimensional space, so $k \ge r$.

We compare two representations of the minimal dimension. Their reachable row spaces and observable column spaces have full dimension: otherwise restriction to reachable rows or quotienting an unobserved direction gives a smaller representation. Map the reachable row after each word in the first representation to the row after that word in the second. Any linear relation in the first vanishes against every suffix. Equality of word weights makes the corresponding relation vanish against all second-representation suffix columns, and full observability makes it zero. The map is well defined. Reversing the argument makes it invertible, with a matrix Ξ satisfying

$$\alpha' = \alpha\Xi, \qquad T'_x = \Xi^{-1} T_x \Xi, \qquad \beta' = \Xi^{-1}\beta. \tag{31}$$

Lemma F.2 supplies basis words of length at most $n-1$. One-letter insertions then use length at most $2n-1$. Dividing (30) by positive history probabilities gives (28). Removing the m^{-k} uniform-action factor in a controlled suffix gives

$$\mathbb{P}_J(o_1, \dots, o_k \mid h, \mathrm{do}(a_1, \dots, a_k)) = m^k \frac{J(hx_1 \cdots x_k)}{J(h)} = m^k v_h T_{x_1} \cdots T_{x_k} \beta. \qquad \square$$

Estimation part of Corollary F.1. We estimate a fixed finite menu and use a vanishing determinant threshold to select the rank. Let $\mathcal{W}$ contain all words of length at most $n-1$. The entries of $\mathsf{H}(\mathcal{W}, \mathcal{W})$ and all one-letter insertions use finitely many cylinders of length at most $2n-1$. We obtain their estimates from Theorem D.1, keeping $J(\varnothing) = 1$ exact. Almost surely all these entries eventually have error at most $3/(2t)$.

At stage t, we choose the largest $k \le n$ for which an estimated k-by-k minor has absolute determinant greater than $t^{-1/2}$, and the first such minor in a fixed ordering. The empty-word entry supplies a one-by-one choice for $t > 1$. Each determinant is a fixed polynomial of finitely many entries in $[0, 1]$, so its error is $O(1/t)$. Every true zero minor eventually falls below the threshold, and every nonzero minor eventually exceeds it. There are finitely many minors. Thus the selected rank is eventually r and the selected basis is eventually fixed. Matrix inversion is continuous at that nonsingular minor, giving

$$\widehat{r}_t = r \text{ eventually}, \qquad (\widehat{\alpha}_t, \widehat{\beta}_t, \widehat{T}_{x,t}) \longrightarrow (\alpha, \beta, T_x) \quad (x \in \Sigma) \quad \text{almost surely.} \qquad \square$$

For every fixed positive-probability history and fixed suffix, the predictions converge by the representation and update equations. The true minor's conditioning affects how late this regime begins. The result estimates a minimal observable representation even when the emission rank is deficient or every hidden action matrix is singular.

F.1 Conditional recovery of hidden parameters

Corollary F.3 (Sufficient conditions for hidden-kernel recovery)**.** *Restrict the model class to full-column-rank emission matrices $E = (\Omega(o \mid s))_{o,s}$ and at least one invertible action kernel. Then J identifies the hidden-state count, E, η_M, and all action kernels up to one common hidden-state permutation.*

Here $E \in \mathbb{R}^{d_o \times n}$ is the column-stochastic emission matrix, of full column rank n. The hidden initial distribution for the controlled tests is a strictly positive column vector η. In the application, it is $\eta_M^\top$. At least one row-stochastic action matrix is invertible. Define the observation marginal μ, the controlled pair matrix Q_b, and the controlled triple tensor $\mathcal{T}_a$ under actions a, a. All three are obtained from J by removing uniform-action factors and marginalizing terminal actions.

The external ingredient is Allman et al. (2009, Corollary 2). For three conditionally independent finite views of an n-class latent variable with strictly positive mixture weights, let their conditional probability matrices have column Kruskal ranks k_1, k_2, k_3. If $k_1 + k_2 + k_3 \ge 2n + 2$, the joint law identifies all three factors and the weights up to one common class permutation. This is the transpose of the source's row convention. Stochastic normalization removes the scaling freedom.

Proof of Corollary F.3. We identify the emissions from three observations, then invert the pair laws to recover every action kernel. Summing over hidden states gives

$$\mu = E\eta, \qquad Q_b = E \,\mathrm{diag}(\eta) P_b E^\top. \tag{32}$$

For a left inverse E^+ of E, multiplying by E^+ and $(E^+)^\top$ shows that the rank of Q_b is the rank of P_b. Thus $n = \max_b \operatorname{rank}(Q_b)$ is determined within the rank-restricted class. We choose an action a of maximal pair rank, so its kernel is invertible.

When $n = 1$, the marginal identifies the sole emission column and all kernels equal $[1]$. For $n \geq 2$, we condition the triple on its middle hidden state. Set

$$\xi = P_a^\top \eta, \qquad L = E \operatorname{diag}(\eta) P_a \operatorname{diag}(\xi)^{-1}, \qquad R = E P_a^\top. \tag{33}$$

An invertible nonnegative matrix has no zero column, so $\xi > 0$. The three observations are conditionally independent given the middle state, with conditional matrices L, E, R. Each is column-stochastic and has full column rank n. The triple tensor is

$$\mathcal{T}_a = \sum_{i=1}^{n} \xi_i \, L(:, i) \otimes E(:, i) \otimes R(:, i).$$

The Kruskal-rank sum is $3n \geq 2n + 2$, so the imported theorem recovers E up to one column permutation. The left view uses ξ, since the initial stationary distribution under averaged actions need not be stationary under action a alone.

We recover the remaining parameters by a left inverse E^+ of the recovered emissions:

$$\eta = E^+ \mu, \qquad P_b = \operatorname{diag}(\eta)^{-1} E^+ Q_b (E^+)^\top \quad (b \in \mathcal{A}). \tag{34}$$

This includes singular kernels for the other actions. For a different recovered labeling $E' = EQ$ with permutation matrix Q, the left inverse $Q^\top E^+$ gives

$$\eta' = Q^\top \eta, \qquad P'_b = Q^\top P_b Q \quad (b \in \mathcal{A}). \qquad \square$$

Within the rank-restricted class, the random-walk law identifies the stationary hidden distribution and the remaining hidden parameters. An arbitrary oracle evaluation prior is a separate parameter. Statistical stability under noisy queries additionally depends on the conditioning of the observable matrices.

G A STATE CAP CHANGES GUARANTEED INFORMATION

For the four-state decoder, C is the largest leaf score and Γ_δ is the gap between the components' possible mean scores. The construction fixes the weights; the supplied loss δ sets the gap, which vanishes linearly at δ_*. We use $0 < \zeta < 1$ for the decoding error probability.

Theorem G.1 (A sharp four-state boundary)**.** *For $0 \leq \delta < 1/2$, recoverable targets are exactly functions of J on the full admissible class with at most two hidden states and arbitrary emissions, and on the class with at most three hidden states and deterministic emissions. In the latter class each equal-law fiber has compatibility diameter at most two.*

For every $29/100 \leq \delta < \delta_*$, where $\delta_* > 293/1000$, an explicit law J_0 with binary observations and four actions has exactly two nonempty compatibility components among all admissible models with at most four states, including unknown emissions. Each has diameter at most two; every such realization has deterministic emissions.

On the promised fiber $J = J_0$, the component bit is decoded from every competent policy with error at most ζ using one native goal, nine prescribed history inputs, and at most $5\lceil 2C^2\Gamma_\delta^{-2} \log(1/\zeta)\rceil$ fresh action samples. Thus four is the smallest deterministic-emission state cap permitting information beyond J for some positive loss, including this interval.

We prove Theorem G.1 through the following finite-state classification: Theorem G.4 settles two states, Theorem G.6 three deterministic-emission states, and Lemma G.7 with Theorems G.12 and G.13 the four-state law. Throughout this appendix $0 \leq \delta < 1/2$ unless a numerical value is specified. Models, native goals, random-walk laws, and statewise competence are those of Section 2; all compared models are communicating and share a possible initial observation. Two models of a class are joined by a compatibility edge when their complete competent-policy sets intersect, as in Section 5; the *compatibility diameter* of a fiber of J is the diameter of this graph on the fiber. The predictive functions $f_{M,s}$ are those of Section 3; the column of a test t is $f_t^M = (f_{M,s}(t))_s$, with $f_\varnothing^M = \mathbf{1}$. For an observation–action pair this appendix writes $\widetilde{B}_{o,a} = D_o P_a$, without the uniform-action factor of (2); that factor is removed from observable cylinders whenever controlled probabilities are used.

G.1 AT MOST TWO HIDDEN STATES

Proposition G.2 (A nondegenerate two-state model is determined up to labels)**.** *Let M have two hidden states with different emission distributions, and suppose $P^M_{a_*}$ is invertible for some action a_*. Every other two-state model N with $J_N = J_M$ is obtained from M by a permutation of hidden-state labels. Admissibility is unnecessary for this proposition.*

Proof. We recover a similarity transformation from finite observable probabilities and use one invertible action to isolate the emission matrices. We choose observations b, c whose emission columns are linearly independent, and an observation d on which the two emission probabilities differ. These choices exist because two different probability distributions are linearly independent. We define the two prefix rows and suffix columns by

$$\Phi_M = \begin{pmatrix} \eta_M D^M_b P^M_{a_*} \\ \eta_M D^M_c P^M_{a_*} \end{pmatrix}, \qquad \Psi_M = \begin{pmatrix} \mathbf{1} & f^M_d \end{pmatrix}. \tag{35}$$

The two entries of η_M are positive by communication. Thus Φ_M and Ψ_M are invertible: the first determinant is the product of the two stationary masses, the selected emission determinant, and $\det P^M_{a_*}$; the second is the difference between the two emission probabilities at d.

We form Φ_N, Ψ_N from the same observable prefixes and suffixes. The entries of $\Phi_M \Psi_M$ are controlled cylinder probabilities, obtained from J_M by removing the known uniform action factors and summing a terminal observation when the suffix is empty. Equality of the laws gives $\Phi_M \Psi_M = \Phi_N \Psi_N$, so both factors for N are invertible. We set $T = \Phi_M^{-1} \Phi_N$ so that $\Phi_N = \Phi_M T$ and $\Psi_N = T^{-1} \Psi_M$. The empty suffix gives $T\mathbf{1} = \mathbf{1}$.

For $\widetilde{B}^M_{o,a} = D^M_o P^M_a$, we insert the same observation-action pair between the selected prefixes and suffixes. Equality of these longer cylinders gives

$$\Phi_M \widetilde{B}^M_{o,a} \Psi_M = \Phi_N \widetilde{B}^N_{o,a} \Psi_N, \qquad \widetilde{B}^N_{o,a} = T^{-1} \widetilde{B}^M_{o,a} T, \qquad P^N_a = T^{-1} P^M_a T. \tag{36}$$

The last identity follows by summing over o; countable sums converge entrywise. The action a_* is invertible in both models. Multiplying (36) by its inverse isolates the emission matrices:

$$D^N_o = T^{-1} D^M_o T \qquad (o \in \mathcal{O}). \tag{37}$$

The diagonal matrix D^M_d has two distinct eigenvalues. Its two eigenspaces are the coordinate axes, so (37) forces each column of T onto a different coordinate axis. The identity $T\mathbf{1} = \mathbf{1}$ sets the two nonzero entries to one. For the resulting permutation matrix T,

$$\Omega_N = T^{-1} \Omega_M, \qquad P^N_a = T^{-1} P^M_a T \qquad (a \in \mathcal{A}). \qquad \square$$

Lemma G.3 (The remaining two-state models admit exact oracles)**.** *A two-state model admits a complete exact statewise native oracle if its two emission distributions coincide or all its action matrices are singular.*

Proof. We first suppose the emission distributions coincide. Every next observation then has the same distribution, independently of the hidden state and chosen action. If all action matrices are singular, their rows coincide because a stochastic two-by-two matrix has determinant zero precisely when its rows are equal. In this case the next observation distribution depends only on the chosen action. In either case, after observing the current observation, the future controlled law has no dependence on the initial hidden state.

We fix a native goal. Its finitely many predicates partition observations into finitely many classes with identical predicate memberships for every action. By Lemma B.2, the pair of the current class and the acceptance monitor is a finite fully observed MDP. Lemma B.2 supplies one policy optimal at every such pair. We use that policy for this goal and choose one for each goal. Since it attains every conditional optimum at the first observation, averaging gives

$$V^{\pi}_{M,s}(g) = V^{*}_{M,s}(g) \qquad (s \in S_M,\ g \in \mathcal{G}). \qquad \square$$

Theorem G.4 (Equal random-walk laws directly share a two-state δ-competent oracle)**.** *On the full class of two-state δ-admissible models, with emissions either unknown or fixed in advance,*

$$J_M = J_N \quad \Longleftrightarrow \quad \mathcal{C}_\delta(M) \cap \mathcal{C}_\delta(N) \neq \varnothing. \tag{38}$$

Consequently the recoverable targets on this class are exactly the functions of its restricted random-walk law J. The same result holds with a state cap of two.

Proof. We assume first that $J_M = J_N$. If either endpoint satisfies Proposition G.2, that proposition makes the endpoints hidden-label permutations, which have identical success values for every observable policy and corresponding hidden states. Their entire competence sets coincide. Otherwise both endpoints satisfy Lemma G.3. They then share an exact oracle by Theorem 3.1, and this oracle is also δ-competent. One-state models have the same observation-only conditional description as Lemma G.3. If a one-state model shares J with a two-state model, the invertible two-by-two cylinder factorization of Proposition G.2 excludes the nondegenerate case, so the same exact-oracle argument applies.

Conversely, a statewise δ-competent oracle is competent under each model's stationary initial distribution by averaging and bounding the supremum of a sum by the sum of suprema. Lemma 3.3 therefore identifies J along every shared edge. Applying these two directions to any decoder and any target F gives

$$F \text{ is recoverable} \quad \Longleftrightarrow \quad F(M) = \tilde{F}(J_M) \text{ on the declared two-state class.} \qquad \square$$

G.2 Deterministic observations with small fibers

We now fix a labeled hidden set S and a deterministic emission map $e : S \to \mathcal{O}$. We write $S_o = e^{-1}(o)$ and $n_o = |S_o|$. Only observations in the finite image of e occur. For a model M with these emissions, let

$$\mathcal{P}_{M,o} = \operatorname{conv}\{f_{M,s} : s \in S_o\}, \qquad \mathcal{Q}_o = \mathcal{P}_{M,o} \cap \mathcal{P}_{N,o}. \tag{39}$$

All convex sets lie in the finite-dimensional span of the two models' state functions, so equality imposes finitely many scalar constraints. A finite generating set means a collection whose convex hull is the given set.

Proposition G.5 (A predictive intersection within the emission budget gives a bridge)**.** *Suppose $J_M = J_N$ and each $\mathcal{Q}_o$ has a generating set of at most n_o points. There is a communicating model R on the same labeled S and the same fixed emission map e such that*

$$J_R = J_M = J_N, \qquad \mathcal{C}_\delta(M) \subseteq \mathcal{C}_\delta(R), \qquad \mathcal{C}_\delta(N) \subseteq \mathcal{C}_\delta(R). \tag{40}$$

Proof. We build one state per common predictive generator, extract a communicating class, and split retained states to restore the prescribed labels.

The generator model. We choose generators $x_{o,1}, \ldots, x_{o,m_o}$ with $m_o \leq n_o$. The sets are nonempty because the stationary predictive function conditional on $O_0 = o$ is the same under the two models and is a mixture of their respective states in S_o.

Every generator $x_{o,i}$ emits o with probability one. After action a, its successor predictive function $y_{o,i,a} = \mathcal{L}_{o,a} x_{o,i}$ is a convex mixture of state functions in both endpoints. We condition this function on its first observation b. When that observation has positive probability, the normalized function is in $\mathcal{Q}_b$: deterministic emissions select only states of S_b in either mixture. We expand these conditional functions in their chosen generators and multiply by their first-observation probabilities. This gives stochastic coefficients $p_{(o,i),a,(b,j)}$ with

$$\mathcal{L}_{o,a} x_{o,i} = \sum_{b,j} p_{(o,i),a,(b,j)} x_{b,j}. \tag{41}$$

We use these coefficients as transitions of a finite model R^0 whose state (o, i) emits o. Induction on test length, using (41) for the step, shows that its state predictive functions are exactly $x_{o,i}$.

Competence transfer. We fix a decomposition $x_{o,i} = \sum_{s \in S_o} u_{(o,i),s} f_{M,s}$. For any observable policy, finite feedback probabilities obey the same mixture identity because the policy's action probabilities depend only on the common observable history. Every native success has a finite witness, so increasing finite acceptance events give the identity for eventual success as well. For every policy π and goal g,

$$V^{\pi}_{R^0,(o,i)}(g) = \sum_s u_{(o,i),s} V^{\pi}_{M,s}(g), \qquad V^{*}_{R^0,(o,i)}(g) \le \sum_s u_{(o,i),s} V^{*}_{M,s}(g). \tag{42}$$

Thus every δ-competent oracle of M is competent in R^0; the same holds for N.

A communicating class. The shared stationary predictive function is a mixture of the $x_{o,i}$, so some initial distribution of R^0 generates J_M. Cesàro averaging its hidden distribution under uniform actions gives a stationary hidden distribution generating the same random-walk law. We decompose the hidden distribution over the closed communicating classes. For each cylinder, Lemma B.1 says that its empirical frequency under J_M converges in probability to the corresponding cylinder probability. Every positive-weight component inherits this convergence, while its stationary expected frequency is its own cylinder probability. Hence every such component has random-walk law J_M. We choose one component and call its restriction R^1. Closedness for the averaged action kernel implies closedness for every action. The competence inequalities (42) survive the restriction.

Restoring the labels. Every observation in the image of e has positive stationary probability in M, because each hidden stationary mass is positive. Therefore R^1 contains at least one state of each such observation and no more than n_o states emitting o. We partition the prescribed set S_o into a nonempty collection of clones for each retained state. We choose positive weights summing to one within each collection, and send every transition into a retained state to its clones with those weights. All clones have the original emission and outgoing transition law. Collapsing clone labels preserves every controlled observable test. Communication holds because each edge into a retained state has positive probability to each of its clones. For each clone s of retained state r and every competent endpoint oracle,

$$V^{\pi}_{R,s}(g) = V^{\pi}_{R^1,r}(g) \ge (1-\delta) V^{*}_{R^1,r}(g) = (1-\delta) V^{*}_{R,s}(g). \qquad \square$$

Theorem G.6 (Emission fibers of size at most two keep each law fiber connected)**.** *Fix S, e with $n_o \le 2$ for every observation. On the full δ-admissible class with these fixed emissions, every J fiber has compatibility diameter at most two. Its recoverable targets are exactly functions of the restricted J.*

In particular this conclusion covers every fixed deterministic emission map on at most three states. It also covers unknown deterministic emissions when the state cap is three.

Proof. Each $\mathcal{P}_{M,o}$ is a point or line segment when $n_o \le 2$. Their nonempty intersection is again a point or line segment and has at most n_o generators; if $n_o = 1$, it is a point. We apply Proposition G.5 to obtain a bridge inside the same class. Every shared edge has equal J by Lemma 3.3, so the two-edge necessity and decoder construction of Theorem 3.6 apply within this fixed class.

With at most three states, an emission fiber of size three means that there is only one possible observation. Every native goal then depends only on chosen actions, so Lemma B.2 gives an exact oracle and Theorem 3.1 gives the conclusion. For unknown deterministic emissions and the state cap three, the endpoints have the same observed labels by equality of J. Except in that one-observation case, every endpoint fiber has size at most two. Their predictive intersection uses no more than the smaller endpoint count for each observation, so the construction of Proposition G.5 uses at most three states. Thus, throughout the declared class,

$$F \text{ is recoverable} \iff F = \tilde{F} \circ J. \qquad \square$$

The generator condition in Proposition G.5 exposes the limit of this argument. In a two-dimensional affine plane, two triangles can intersect in a hexagon. Consequently an emission fiber with three states need not have an intersection with at most three generators. The construction below realizes this obstruction as a disconnected compatibility fiber.

G.3 A FOUR-STATE LAW WITH TWO COMPONENTS

We construct a random-walk law whose four-state admissible realizations form two compatibility components. Exhaustive realization classification is needed to exclude a connection through a third four-state model.

G.3.1 THE CONTROLLED MODEL AND ITS REALIZATIONS

The hidden states are $1, 2, 3, *$, with deterministic observations $0, 0, 0, 1$. There are four actions $\mathsf{r}, \mathsf{s}, \mathsf{t}, \mathsf{z}$. We use the row vectors and matrices

$$h = (1/3, 1/3, 1/3), \qquad x = (0, 1/3, 2/3), \qquad H = \mathbf{1}h, \qquad \varepsilon = 19/100,$$
$$\Pi_{\mathrm{cyc}} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}, \qquad \Pi_{\mathrm{swp}} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \qquad p = \tfrac{1}{2}\mathbf{1} + \varepsilon(-1, 0, 1)^{\mathsf{T}}. \tag{43}$$

The base model M_1 has transition matrices

$$P_{\mathsf{r}} = \begin{pmatrix} \Pi_{\mathrm{cyc}} & 0 \\ x & 0 \end{pmatrix}, \qquad P_{\mathsf{s}} = \begin{pmatrix} \Pi_{\mathrm{swp}} & 0 \\ x & 0 \end{pmatrix}, \qquad P_{\mathsf{t}} = \begin{pmatrix} (\mathbf{1} - p)x & p \\ x & 0 \end{pmatrix}, \qquad P_{\mathsf{z}} = \begin{pmatrix} \mathbf{1}x & 0 \\ x & 0 \end{pmatrix}. \tag{44}$$

For a nonzero real number τ, let

$$A_\tau = \mathrm{diag}\big(H + \tau(I - H), 1\big), \quad P_a^{(\tau)} = A_\tau P_a A_\tau^{-1}, \quad x_\tau = h + \frac{x - h}{\tau}, \quad p_\tau = \tfrac{1}{2}\mathbf{1} + \tau\varepsilon(-1, 0, 1)^{\mathsf{T}}. \tag{45}$$

The matrices in (44) with x, p replaced by x_τ, p_τ give precisely $P_a^{(\tau)}$. We call this model M_τ whenever its entries are nonnegative. The law J_0 is the random-walk law of M_1.

Figure 2 uses the tests t and rt, each accepting a final observation one, as coordinates. They determine the zero-emitting state functions affinely.

Lemma G.7 (All realizations with at most four states lie on two scalar intervals)**.** *Every finite state-emitting model with at most four hidden states and random-walk law J_0 has exactly four states. Up to hidden-state permutation it is M_τ for a unique*

$$\tau \in [-50/19, -1] \cup [1, 50/19]. \tag{46}$$

In particular its emissions are necessarily deterministic with multiplicities three and one, even if the model class does not specify emissions in advance. Every model in (46) *is communicating and has random-walk law J_0.*

Proof. We first recover the entire predictive linear representation from four independent observable tests, then impose nonnegativity on its hidden realization.

A four-dimensional representation. We use the test columns

$$\Psi = \big[\mathbf{1},\ D_1\mathbf{1},\ D_0P_{\mathsf{t}}D_1\mathbf{1},\ D_0P_{\mathsf{r}}D_0P_{\mathsf{t}}D_1\mathbf{1}\big]. \tag{47}$$

Their determinant is $-3\varepsilon^2$. We choose prefixes ending just after an action, before the next emission, so that every suffix in (47) starts at a new observation. After removing the known uniform-action factors, their unnormalized state rows are

$$\Phi = \begin{bmatrix} \eta D_1 P_{\mathsf{z}} \\ \eta D_1 P_{\mathsf{z}} D_0 P_{\mathsf{r}} \\ \eta D_1 P_{\mathsf{z}} D_0 P_{\mathsf{r}} D_0 P_{\mathsf{r}} \\ \eta D_0 P_{\mathsf{t}} \end{bmatrix}.$$

The first three rows are $\eta_*(x, 0)$, $\eta_*(x\Pi_{\mathrm{cyc}}, 0)$, and $\eta_*(x\Pi_{\mathrm{cyc}}^2, 0)$. The fourth has last coordinate $\eta_0 p > 0$, where $\eta_0 = (\eta_1, \eta_2, \eta_3)$, because stationarity is strictly positive and every entry of p is positive. Since the three cyclic rows of x have determinant $1/3$, we obtain

$$\det \Phi = \frac{\eta_*^3 \eta_0 p}{3} > 0, \qquad \det(\Phi\Psi) = -\eta_*^3(\eta_0 p)\varepsilon^2 \neq 0.$$

Each entry of $\Phi\Psi$ is an observable cylinder probability with its prescribed action factors removed. The empty suffix supplies a prefix marginal. This factorization also holds in a stochastic-emission realization N: the rows use the same action-ending prefixes and the columns the same suffixes, with no observation repeated at the joining time. Inserting one additional pair (o, a) at the join similarly gives the observable matrix $\Phi D_o P_a \Psi$. Equality of J makes both these cylinder matrices agree for M and N.

Exactly four states. Every realization factors $\Phi\Psi$ through its hidden dimension. Thus a realization with at most four states has exactly four, with invertible prefix and suffix factors. We put $A = \Psi_N \Psi_M^{-1}$. The empty-suffix column gives $A\mathbf{1} = \mathbf{1}$, and $\Phi_M \Psi_M = \Phi_N \Psi_N$ gives $\Phi_M = \Phi_N A$. The inserted-pair identities then give

$$\widetilde{B}_{o,a}^N = A\widetilde{B}_{o,a}^M A^{-1}, \qquad P_a^N = A P_a^M A^{-1}, \qquad \widetilde{B}_{o,a} = D_o P_a. \tag{48}$$

The emissions. We can isolate the emissions without an invertible individual action. The matrix consisting of the first three columns of P_{r} and the last column of P_{t} has determinant $(2\varepsilon - 3)/6 \neq 0$. Hence the column spaces of the action matrices jointly span all four coordinates. From (48),

$$\big(A^{-1} D_o^N A - D_o^M\big) P_a^M = 0 \quad \text{for every } a, \qquad D_o^N = A D_o^M A^{-1}. \tag{49}$$

Since emission matrices are diagonal, their diagonal entries are their eigenvalues. The spectrum of D_1^M is $(0, 0, 0, 1)$, the spectrum of D_0^M is $(1, 1, 1, 0)$, and all other emission matrices vanish. We relabel N so its emissions agree with those of M. Equation (49) then makes A block diagonal, with a three-by-three block on the zero-emitting states and scalar one on $*$.

The zero-emitting triangle. We identify the predictive function of each zero-emitting state with its coefficient row in the three base zero-emitting state functions. These rows sum to one. They are affinely independent because A is invertible, so their convex hull is a triangle in the plane of coordinate sum one. Under r and s, the next observation from this triangle is certainly zero, as follows from its predictive functions. Nonnegative transitions therefore map this triangle into itself by the permutations Π_{cyc} and Π_{swp}. Since $\Pi_{\mathrm{cyc}}^3 = \Pi_{\mathrm{swp}}^2 = I$, these inclusions are equalities.

The two permutations generate the full coordinate-permutation group on three coordinates. They permute the three vertices faithfully: an element fixing all vertices fixes their affine span and is the identity. We label the vertices so that the cycle and transposition act on their labels as on the coordinate labels. The vertex labeled i is then fixed by the transposition of the other two coordinates, so those two coordinates coincide. Its coordinate sum is one, and equivariance gives one common scalar $\tau \neq 0$:

$$\text{the three predictive rows are } h + \tau(e_i - h), \qquad i = 1, 2, 3. \tag{50}$$

Equation (50) gives $A = A_\tau$ after the chosen label permutation. The row leaving $*$ under z is x_τ. Its nonnegativity requires $|\tau| \geq 1$. The three probabilities in p_τ require $|\tau|\varepsilon \leq 1/2$. All remaining entries in (44)-(45) are products of these nonnegative probabilities, giving exactly (46).

The converse. Each parameter in (46) makes all four matrices stochastic. Rotation connects the three zero-emitting states; the middle coordinate has test probability $1/2$ to reach $*$; and $*$ returns to a nonempty zero-emitting support. Thus each model communicates. Similarity carries the stationary vector of M_1 to a real stationary vector of total mass one in M_τ. Uniqueness of the stationary vector in this communicating stochastic model makes it its positive stationary distribution. Since the similarity commutes with emissions, every stationary cylinder satisfies

$$J_{M_\tau}(o_0 a_0 \cdots o_k) = J_{M_1}(o_0 a_0 \cdots o_k) = J_0(o_0 a_0 \cdots o_k). \qquad \square$$

G.3.2 A UNIFORM COMPETENCE BOUND FOR THE COMPLETE NATIVE LANGUAGE

Let $\mathcal{V}_0$ be the six permutations of $(\ell, m, u) = (31/100, 1/2, 69/100)$, viewed as nonnegative value vectors for the three zero-emitting initial states. Let $\operatorname{cone} \mathcal{V}_0$ be their nonnegative linear span. For a nonzero nonnegative vector w, define

$$\mu_i(w) = \min_{\theta \in \mathcal{V}_0} \frac{w \cdot \theta}{\theta_i}, \qquad \kappa(w) = \sum_{i=1}^{3} \frac{w_i}{\mu_i(w)}. \tag{51}$$

All denominators are positive because every entry of every generator is positive.

Lemma G.8 (A cone comparison)**.** *For every nonzero $w \geq 0$, $\kappa(w) \leq 100/71$. Consequently every nonempty compact convex set $\mathcal{H} \subseteq \operatorname{cone} \mathcal{V}_0$ contains a vector v satisfying*

$$v_i \geq \frac{71}{100} \max_{q \in \mathcal{H}} q_i \qquad (i = 1, 2, 3). \tag{52}$$

Proof. We normalize w to sum one and order its coordinates as $a \geq b \geq c \geq 0$. To maximize $\theta_i/(w \cdot \theta)$, we put the largest generator coordinate u at i: exchanging it with a smaller coordinate increases that ratio. Among the remaining coordinates, placing ℓ at the larger weight minimizes the denominator. Thus

$$\kappa(w) = u \left(\frac{a}{ua + \ell b + mc} + \frac{b}{\ell a + ub + mc} + \frac{c}{\ell a + mb + uc} \right). \tag{53}$$

We denote the three denominators by d_1, d_2, d_3. They are at least ℓ. The inequality $\kappa(w) \leq 100/71$ is equivalent to the nonnegativity of

$$P(a, b, c) = 100 d_1 d_2 d_3 - 71u\big(a d_2 d_3 + b d_1 d_3 + c d_1 d_2\big). \tag{54}$$

We set $a = x + y/2 + z/3$, $b = y/2 + z/3$, and $c = z/3$, so $x, y, z \geq 0$ and $x + y + z = 1$. As shown in Appendix G.5, $(x + y + z)^{22} P(a, b, c)$ has nonnegative coefficients, so $P \geq 0$ and $\kappa(w) \leq 100/71$.

We now put $\bar{q}_i = \max_{q \in \mathcal{H}} q_i$. For every $w \geq 0$, a generator decomposition of a vector $q \in \operatorname{cone} \mathcal{V}_0$ gives $w \cdot q \geq \mu_i(w) q_i$. Choosing a vector attaining each $\bar{q}_i$ therefore gives

$$\max_{q \in \mathcal{H}} w \cdot q \geq \max_i \mu_i(w) \bar{q}_i, \qquad w \cdot \bar{q} \leq \kappa(w) \max_i \mu_i(w) \bar{q}_i \leq \frac{100}{71} \max_{q \in \mathcal{H}} w \cdot q. \tag{55}$$

We now show that these inequalities force $\mathcal{H}$ to meet the upper orthant $Q = (71/100)\bar{q} + \mathbb{R}^3_+$. If the two convex sets were disjoint, we would choose a closest pair $v \in \mathcal{H}, q \in Q$, which exists because $\mathcal{H}$ is compact. Projection onto the orthant gives $w = q - v \geq 0$, $w \neq 0$, and $w \cdot q = (71/100) w \cdot \bar{q}$.

Minimality of the distance, applied to segments in $\mathcal{H}$, gives $w \cdot v' \le w \cdot v$ for all $v' \in \mathcal{H}$. This contradicts (55), since

$$\max_{v' \in \mathcal{H}} w \cdot v' \le w \cdot v < w \cdot q = \frac{71}{100} w \cdot \bar{q}. \qquad \square$$

Lemma G.9 (Both orientations are fully δ-admissible at $\delta = 29/100$)**.** *The models M_1 and M_{-1} admit complete statewise native oracles at $\delta = 29/100$.*

Proof. We reduce every native goal to a finite menu of value vectors before the first action that erases the initial hidden coordinate. We fix its finite acceptance monitor from Lemma B.2 and first suppose the initial observation is zero. Until the first t or z, actions r, s only permute the hidden coordinate and all observations remain zero. The current permutation and monitor state form a finite, fully observed control state.

Beliefs after the first test or reset. After z, the hidden law is x_τ; after t, it is either x_τ or the known state $*$, depending on the observed outcome. These laws are independent of the hidden state before that action, conditional on the outcome. Subsequent beliefs are among the six permutations of x_τ and the point mass at $*$. The belief and monitor therefore form a finite fully observed MDP. By Lemma B.2 we choose continuations optimal at every such pair. This choice is simultaneously optimal for all three original hidden contexts once their histories reach the same observed outcome.

Exit vectors. Before this first t or z, each reachable exit from the finite permutation-monitor graph now has a fixed three-context value vector. An immediate success or failure gives $\mathbf{1}$ or zero. A reset exit gives a constant vector. A test exit, for some permutation ι, gives

$$z_0\big(\mathbf{1} - p_{\tau,\iota}\big) + z_1 p_{\tau,\iota}, \qquad z_0, z_1 \in [0,1]. \tag{56}$$

For $\tau = 1$ or -1, both $p_{\tau,\iota}$ and its complement belong to $\mathcal{V}_0$. Constants belong to $\operatorname{cone} \mathcal{V}_0$ because averaging its six generators gives $\mathbf{1}/2$. Thus every exit vector lies in $\operatorname{cone} \mathcal{V}_0$.

The attainable hull. There are finitely many exit vectors. Every reachable exit is attainable by a finite word in r, s before its final exit action. A policy before the first exit induces a distribution on this finite list; any probability of never exiting and never accepting contributes zero. Conversely, a mixture of finitely many chosen exit words is implemented by behavioral action probabilities along their common prefixes. Hence the convex hull $\mathcal{H}$ of the reachable exit vectors and zero is compact. Every point of this hull is dominated coordinatewise by an attainable point: if zero is unattainable, we assign its mixture mass to any attainable exit vector. Replacing any later continuation by the selected optimal one can only improve all initial contexts, so the coordinate maxima of $\mathcal{H}$ are exactly their native goal benchmarks.

The policy. Lemma G.8 supplies a vector at least $71/100$ of all three benchmarks, and we choose an attainable dominating vector for this initial observation. For initial observation one, the hidden state is $*$ and the finite belief-monitor MDP has an optimal policy by Lemma B.2. These two choices define one policy for the goal; we make such a choice for every goal. For both endpoint orientations,

$$V^{\pi}_{M_{\pm 1},s}(g) \ge \frac{71}{100} V^{*}_{M_{\pm 1},s}(g) \qquad (s \in \{1,2,3,*\},\ g \in \mathcal{G}). \qquad \square$$

G.3.3 One explicit native goal separates the orientations

After a reset z, a test t has success probability

$$q = xp = 47/75. \tag{57}$$

Repeating reset and test produces independent Bernoulli coins with this probability in both endpoint models. We use the prefix-free event

$$L = \{0, 10, 110, 1110, 111100, 1111010, \\ 11110110, 111101110, 1111011110, 11110111110\}. \tag{58}$$

on these successive test outcomes. It is the lexicographic event that a twelve-bit coin string has integer value less than 3966. Its probability β satisfies

$$899/1000 \le \beta \le 9/10. \tag{59}$$

Each prefix in (58) is a finite current/next goal with prescribed reset and test actions. Their disjunction is native. Appendix G.5 relates the probability bounds to this event; the exact evaluation is in the repository.

We define one native goal g_* by three branches. All require initial observation zero, a specified four-action word in r, s while observations remain zero, then action t and next observation one. The words are

$$w_1 = \mathsf{rrss}, \qquad w_2 = \mathsf{rrsr}, \qquad w_3 = \mathsf{rsrs}. \tag{60}$$

Branches one and two additionally require a reset and the lottery (58). Branch three accepts at its first successful test. The three words are distinct and have equal length, so their branch choices occur before any informative observation. The ideal continuation value matrices, with one row per branch and one column per zero-emitting initial state, are

$$B_+ = \begin{pmatrix} \beta u & \beta \ell & \beta m \\ \beta \ell & \beta u & \beta m \\ \ell & m & u \end{pmatrix}, \qquad B_- = \begin{pmatrix} \beta \ell & \beta u & \beta m \\ \beta u & \beta \ell & \beta m \\ u & m & \ell \end{pmatrix}. \tag{61}$$

Here B_+ belongs to M_1, and B_- to M_{-1}.

Lemma G.10 (The native goal has no common competent oracle)**.** *No policy is* $29/100$*-competent for* g_* *in both* M_1 *and* M_{-1}*.*

Proof. We let v_1, v_2, v_3 be the probabilities that the policy follows the respective four-action word and then the required test. Since all preceding observations are zero, these are the same in both models, nonnegative, and sum to at most one. Failing to follow the prescribed lottery actions can only decrease the branch value, so competence implies the corresponding inequalities from the ideal matrices (61).

We use the first two initial states of M_1 and the first initial state of M_{-1}. Their benchmarks are $\beta u, \beta u, u$, respectively, by (59). We multiply these three necessary competence inequalities by $19, 10, 9$, chosen so that v_1 and v_2 get equal coefficients (recall $\ell + u = 1$). The resulting coefficients of (v_1, v_2, v_3) are $(19\beta, 19\beta, 171/10)$. Since $\beta \le 9/10$, their weighted sum is at most $171/10$. By (59), the weighted benchmark exceeds this upper bound, giving

$$\frac{171}{10} \ge 19\beta(v_1 + v_2) + \frac{171}{10} v_3 \ge \frac{71}{100} u(29\beta + 9) > \frac{171}{10}. \qquad \square$$

G.3.4 The complete compatibility classification on this fiber

Theorem G.11 (Exactly two components, with a recoverable bit beyond J)**.** *Fix* $\delta = 29/100$*, the four actions in* (44)*, binary observations, and a state cap of four. Consider every communicating, statewise* δ*-admissible model with random-walk law* J_0*, without fixing its emissions. The compatibility graph has exactly two nonempty components:*

$$\{M_\tau : \tau \ge 1,\ \mathcal{C}_\delta(M_\tau) \ne \varnothing\}, \qquad \{M_\tau : \tau \le -1,\ \mathcal{C}_\delta(M_\tau) \ne \varnothing\}, \tag{62}$$

where hidden-state permutations represent the same displayed model. Each component has diameter at most two. The sign of τ *is recoverable from every competent complete oracle on this fiber, although*

it is not a function of J. The theorem also holds when the four labeled states and deterministic emissions $(0, 0, 0, 1)$ are fixed in advance.

Proof. We use the entire realization classification to exclude paths through other models. Lemma G.7 places every model with at most four states and law J_0 in (46). For $1 \le \tau' \le \tau$, each zero-emitting predictive state of $M_{\tau'}$ is the stochastic mixture of the corresponding states of M_τ with matrix

$$U = H + \frac{\tau'}{\tau}(I - H). \tag{63}$$

The state $*$ has the same predictive function in all models. As in (42), mixtures preserve every observable policy's success probability and bound the mixture benchmark by the mixture of benchmarks. Hence

$$\mathcal{C}_\delta(M_\tau) \subseteq \mathcal{C}_\delta(M_1) \quad (\tau \ge 1), \qquad \mathcal{C}_\delta(M_\tau) \subseteq \mathcal{C}_\delta(M_{-1}) \quad (\tau \le -1). \tag{64}$$

Lemma G.9 makes the two central endpoints admissible. Every admissible positive parameter therefore has a shared edge with M_1, and every admissible negative parameter has one with M_{-1}. Lemma G.10 and (64) exclude every edge between the two parameter signs. Lemma 3.3 excludes edges from this fiber to a model with another random-walk law, giving (62).

For an explicit decoder on this fiber, we query the policy for g_* at the histories along the three words in (60), and at their test choices. Their products give (v_1, v_2, v_3) from Lemma G.10. Let Q_+ and Q_- be the sets of nonnegative triples of total mass at most one satisfying all three necessary competence inequalities from B_+ and B_-, respectively. Lemma G.10 proves $Q_+ \cap Q_- = \varnothing$. By (64), a competent oracle of a positive parameter has its triple in Q_+, while a competent oracle of a negative parameter has it in Q_-. Thus the finite-query rule

$$D(\pi) = \begin{cases} +1, & (v_1, v_2, v_3) \in Q_+, \\ -1, & \text{otherwise}, \end{cases} \qquad D(\pi) = \operatorname{sign}(\tau) \quad \big(\pi \in \mathcal{C}_\delta(M_\tau)\big). \qquad \square$$

The decoder needs at most fifteen point queries to action distributions, five along each of the three branches, with repetitions removed if desired. It is conditional on the promised fiber J_0. On the whole four-state δ-admissible class, define a target equal to this sign on J_0 and zero on every other fiber of J. Lemma 3.3 first determines J from the complete oracle, and then the finite decoder above recovers this target. Its values at M_1 and M_{-1} differ despite their equal J.

This counterexample does not conflict with Theorem 3.6 on the class of all finite models. In that larger class, the six-vertex predictive hexagon can be realized with six zero-emitting states and one one-emitting state; the four-state restriction excludes such a bridge.

G.4 A LOSS INTERVAL AND A SAMPLED COMPONENT DECODER

The strict separation in Lemma G.10 gives a loss interval and a statistical decoder. We use the native goal g_* in (60), with lottery probability β satisfying (59). The score gap vanishes at

$$\delta_* = 1 - \frac{171}{10u(29\beta + 9)}, \qquad u = 69/100. \tag{65}$$

We combine the first two columns of B_+ to form the separating score:

$$\begin{gathered} \omega = B_+(19, 10, 0)^{\mathsf{T}}, \qquad C = \omega_1, \\ A_\delta = (1 - \delta)29\beta u, \qquad B_\delta = \frac{171}{10} - (1 - \delta)9u, \\ \Gamma_\delta = A_\delta - B_\delta = u(29\beta + 9)(\delta_* - \delta). \end{gathered} \tag{66}$$

For a policy π, let v_i be its probability of following the ith four-action word in (60) and then choosing t, when every input observation before that final action is zero. Thus v_i excludes the test outcome and any subsequent lottery, and $v_1 + v_2 + v_3 \le 1$.

Theorem G.12 (Two components throughout an explicit loss interval)**.** *For every* $29/100 \le \delta < \delta_*$*, the full admissible* J_0 *fiber with at most four hidden states has exactly the two components in* (62)*, each of diameter at most two. For every competent policy,*

$$\omega \cdot v \ge A_\delta \quad \textit{in the positive component}, \qquad \omega \cdot v \le B_\delta \quad \textit{in the negative component}. \tag{67}$$

Here $\Gamma_\delta > 0$*. Moreover* $\delta_* > 293/1000$*.*

Proof. We keep the two endpoint guarantees separate in the weighted separation argument. The necessary competence inequalities from (61) hold at any loss δ. In M_1, the first two benchmarks are $\beta u, \beta u$, where $u = 69/100$. Multiplying these inequalities by 19 and 10 gives $\omega \cdot v \ge A_\delta$. In M_{-1}, the first benchmark is u, so the first column of B_-, denoted by b_-, satisfies $9b_- \cdot v \ge (1-\delta)9u$. Directly from the three rows in (61),

$$\omega + 9b_- = (19\beta, 19\beta, 171/10).$$

Since $\beta \le 9/10$ and $\sum_i v_i \le 1$, we obtain $\omega \cdot v \le 171/10 - (1-\delta)9u = B_\delta$ in M_{-1}. Equation (65) therefore gives the positive gap Γ_δ for $\delta < \delta_*$.

We extend these endpoint bounds to every realization in the fiber. Lemma G.7 classifies all such models, including unknown emissions, as M_τ in (46). For $|\tau| \ge 1$, the predictive states of $M_{\mathrm{sign}(\tau)}$ are mixtures of those of M_τ, with matrix $H + |\tau|^{-1}(I - H)$ on the zero-emitting states and the identity on $*$. This matrix is stochastic. For any native goal, its mixture identity preserves each policy's value and bounds the mixture optimum by the mixture of statewise optima, exactly as in (42). Consequently, for the present loss δ,

$$\mathcal{C}_\delta(M_\tau) \subseteq \mathcal{C}_\delta(M_{\mathrm{sign}(\tau)}).$$

Both endpoints remain admissible for $\delta \ge 29/100$ by Lemma G.9. Every admissible realization therefore shares an edge with its endpoint. The positive score gap excludes every edge between the two signs, and Lemma 3.3 excludes paths leaving the fiber of J. Thus each sign is one nonempty component of diameter at most two. The lottery evaluation in Appendix G.5 gives

$$\delta_* > \frac{293}{1000} > \frac{29}{100}. \qquad \square$$

Theorem G.13 (Finite-confidence decoding from sampled actions)**.** *Fix a known* $29/100 \le \delta < \delta_*$ *and* $0 < \zeta < 1$*. The unknown model is promised to have at most four states and law* J_0*. Let* π *be any complete* δ*-competent policy, accessible through fresh independent action samples at prescribed history–goal inputs. Its component is recovered with probability at least* $1 - \zeta$ *using only* g_**, nine distinct history inputs, and at most* $5K$ *action samples, where*

$$K = \left\lceil \frac{2C^2}{\Gamma_\delta^2} \log \frac{1}{\zeta} \right\rceil. \tag{68}$$

The probability is over the policy-query samples, uniformly over every model and competent policy satisfying the promise. The decoder uses no environment trajectories or exact action probabilities.

Proof. We sample the policy along a finite tree whose leaf scores have expectation $\omega \cdot v$. Its three accepted action words are

$$w_1\mathsf{t} = \mathsf{rrsst}, \qquad w_2\mathsf{t} = \mathsf{rrsrt}, \qquad w_3\mathsf{t} = \mathsf{rsrst}.$$

We start a replicate by querying $\pi(\cdot \mid 0, g_*)$. After each sampled r or s, we append that action and a supplied zero observation to the next query input. If the sampled action leaves the prefix tree, we stop the replicate and assign score zero. At a four-action prefix we sample the fifth action; the score is ω_i precisely when the result is the full word $w_i\mathsf{t}$, and is zero otherwise. The distinct action prefixes at which a query can occur are

$$\varnothing,\ \mathsf{r},\ \mathsf{rr},\ \mathsf{rs},\ \mathsf{rrs},\ \mathsf{rsr},\ \mathsf{rrsr},\ \mathsf{rrss},\ \mathsf{rsrs}.$$

Each prefix specifies one alternating history with all observations zero. These are policy-function inputs; appending zero does not assert that an environment trajectory has been observed.

Because the draws are conditionally fresh, the probability of completing $w_i\mathsf{t}$ is the product of its five action probabilities, exactly v_i. Since $0 \le \omega_i \le C$, one replicate has score $X \in [0, C]$ with

$\mathbb{E}X = \omega \cdot v$. We repeat the procedure K times. Fresh independent randomness at every call makes the replicate scores independent and identically distributed, even though each replicate chooses its next input from earlier replies.

We declare the positive component when the average score is at least $(A_\delta + B_\delta)/2$, and the negative component otherwise. By Theorem G.12, either error requires a one-sided deviation of at least $\Gamma_\delta/2$ from the mean. Hoeffding's bounded independent-variable inequality gives

$$\mathbb{P}(\text{incorrect component}) \le \exp\left(-\frac{K\Gamma_\delta^2}{2C^2}\right) \le \zeta. \qquad \square$$

For fixed confidence, the sample bound grows as $(\delta_* - \delta)^{-2}$. With exact action distributions, the same nine history queries determine the branch probabilities and their score directly. Neither version tests the promise $J = J_0$ with finitely many queries.

G.5 Finite certificates for the four-state construction

The cone inequality has a finite coefficient certificate. Write the cubic in (54), after the ordered-simplex substitution, as

$$P = \sum_{u+v+w=3} p_{uvw} x^u y^v z^w. \tag{69}$$

Multiplication by $(x+y+z)^{22}$ gives

$$(x+y+z)^{22} P = \sum_{i+j+k=25} c_{ijk} x^i y^j z^k, \qquad c_{ijk} = \sum_{u+v+w=3} p_{uvw} \binom{22}{i-u, j-v, k-w},$$

where a multinomial coefficient with a negative argument is zero. All coefficients are nonnegative. Since $x+y+z=1$ on the simplex, this proves the inequality used in Lemma G.8. The coefficient list and exact expansion are supplied in the repository.

For the native lottery, let $n_1(\ell)$ and $n_0(\ell)$ count the ones and zeros in a prefix $\ell \in L$, with L from (58) and $q = 47/75$. Disjointness of the prefixes gives

$$\beta = \sum_{\ell \in L} q^{n_1(\ell)} (1-q)^{n_0(\ell)}.$$

Summing the disjoint prefix probabilities gives (59). Since δ_* in (65) increases with β, the lower bound in (59) gives $\delta_* > 293/1000$. The repository also supplies the exact sum, the matrix identities for the three branch words and the action-ending-prefix factorization.

G.6 Stochastic emissions and the state cap

We retain the model and competence definitions of Section 2, with $0 \le \delta < 1/2$. An n-state realization is *minimal* when its Hankel matrix of J has rank n. Its action kernels have *full joint image* when

$$\operatorname{rank}[P_{a_1} \ \cdots \ P_{a_m}] = n. \tag{70}$$

One invertible action mixture suffices for this condition. An *emission profile* is the complete distribution $\Omega(\cdot \mid s)$; states with the same profile form a block.

Lemma G.14 (Stationary similarity preserves emission profiles)**.** *Let M be a minimal n-state model satisfying* (70)*. Every realization N of J_M with at most n states has exactly n states, and an invertible matrix T satisfies*

$$\begin{aligned} T\mathbf{1} &= \mathbf{1}, \qquad \eta_N = \eta_M T, \\ P_a^N &= T^{-1} P_a^M T, \qquad D_o^N = T^{-1} D_o^M T. \end{aligned} \tag{71}$$

The two models have the same profiles and multiplicities. After matching their blocks, T is block diagonal.

Proof. We choose a nonsingular n-by-n Hankel minor and its prefix–suffix factorizations $\Phi_M \Psi_M = \Phi_N \Psi_N$. Rank forces N to have n states and all four factors to be invertible. We set $T = \Phi_M^{-1} \Phi_N$, so

that $\Psi_N = T^{-1}\Psi_M$. Equality at the empty suffix gives $T\mathbf{1} = \mathbf{1}$; equality of the probabilities under J of the selected suffixes gives $\eta_N = \eta_M T$.

Inserting an observation-action pair between these prefixes and suffixes gives $D_o^M P_a^M T = T D_o^N P_a^N$. Summing over o gives $P_a^M T = T P_a^N$, and substitution yields

$$(D_o^M T - T D_o^N) P_a^N = 0 \qquad (a \in \mathcal{A}).$$

The joint-image rank is preserved by similarity. Since the columns of the P_a^N jointly span $\mathbb{R}^n$, the first factor vanishes. Entrywise,

$$\big(\Omega_M(o \mid i) - \Omega_N(o \mid j)\big) T_{ij} = 0 \qquad (o \in \mathcal{O}).$$

Every nonzero entry joins identical profiles. Invertibility makes each matched block square and nonsingular, so

$$|S_\lambda^M| = |S_\lambda^N|, \qquad T = \operatorname{diag}(T_\lambda)_\lambda. \qquad \square$$

Theorem G.15 (Two states per stochastic emission profile)**.** *Fix n. Consider the communicating, minimal, δ-admissible n-state models satisfying* (70)*, with at most two states per emission profile. If $J_M = J_N$, a model R in this same class satisfies*

$$J_R = J_M = J_N, \qquad \mathcal{C}_\delta(M) \cup \mathcal{C}_\delta(N) \subseteq \mathcal{C}_\delta(R). \tag{72}$$

Every fiber of J has compatibility diameter at most two, and the recoverable targets on this class are exactly functions of J.

Proof. We intersect the endpoint state mixtures within each profile, where each intersection is a point or line segment. By Lemma G.14, we may match the blocks and use its block diagonal T. For a block I, let Δ_I denote the probability row vectors supported on I, and put

$$Q_I = \{u \in \Delta_I : uT \in \Delta_I\}. \tag{73}$$

Both stationary distributions are positive. Because $\eta_N = \eta_M T$ and $T_I \mathbf{1} = \mathbf{1}$, their common block mass $p_I > 0$ gives a relative interior point $\eta_M|_I / p_I$ of Q_I. Thus Q_I is a point for $|I| = 1$ and a nondegenerate interval for $|I| = 2$.

The bridge. We choose its endpoints as the rows of a block diagonal matrix U_M. The matrices U_M and $U_N = U_M T$ are invertible and stochastic. Each row of U_M becomes one state of R, with its block's emission profile. For a generator u and action a, the successor row $y = uP_a^M$ is a probability vector, and so is $yT = uTP_a^N$. Their restrictions to a block have equal mass. Each positive-mass restriction, normalized, belongs to $Q_{I'}$ for its own block I'. Expressing these restrictions in the chosen endpoints of each block defines stochastic transitions with

$$P_a^R U_M = U_M P_a^M, \qquad P_a^R U_N = U_N P_a^N, \qquad D_o^R U_M = U_M D_o^M. \tag{74}$$

Communication. The interior stationary points give positive barycentric weights η_R satisfying $\eta_R U_M = \eta_M$. Equation (74) makes η_R stationary for $\overline{P}^R$. This matrix is similar to the irreducible $\overline{P}^M$, so its eigenvalue one is simple. Positivity of a stationary distribution excludes transient states; multiple closed classes would give multiple stationary distributions. Hence R communicates.

Competence and the class. Induction on finite tests gives the state predictive mixtures specified by U_M. Multiplying by the policy's history-dependent action probabilities extends this identity to finite feedback events, and increasing finite acceptance events gives

$$V_{R,i}^\pi(g) = \sum_s (U_M)_{is} V_{M,s}^\pi(g), \qquad V_{R,i}^*(g) \le \sum_s (U_M)_{is} V_{M,s}^*(g). \tag{75}$$

The same calculation uses U_N for N, proving the competence inclusions. The stationary mixture has law J_M, so R remains minimal; similarity preserves (70), and every profile keeps its multiplicity. Since $\mathcal{C}_\delta(R)$ contains the nonempty set $\mathcal{C}_\delta(M)$, the model R is δ-admissible and belongs to the class. Both endpoints are admissible, so M–R and R–N are edges, and the fiber has diameter at most two.

Recoverable targets. If a decoder D returns F on every competent oracle of the class, we choose $\sigma_M \in \mathcal{C}_\delta(M)$ and $\sigma_N \in \mathcal{C}_\delta(N)$. Both lie in $\mathcal{C}_\delta(R)$, so $F(M) = D(\sigma_M) = F(R) = D(\sigma_N) = F(N)$, and F is constant on the fiber. Conversely, a shared competent oracle determines J by Lemma 3.3, so $D(\pi) = \tilde{F}(J_M)$ for $\pi \in \mathcal{C}_\delta(M)$ is well defined. Together,

$$F \text{ is recoverable on this class} \quad \Longleftrightarrow \quad F = \tilde{F} \circ J. \qquad \square$$

Corollary G.16 (Three states with stochastic emissions)**.** *On the class of communicating, δ-admissible models with at most three states, Hankel rank three for J, and full joint action image, every J fiber has compatibility diameter at most two. Its recoverable targets are exactly functions of J. The emissions may be unknown and have rank two.*

Proof. We compare two models in the class. Hankel rank three forces both to have three states. A single common emission profile would make observations state-independent under every action sequence, giving Hankel rank one. Every profile therefore occurs at most twice. Applying Theorem G.15 gives

$$F \text{ is recoverable} \quad \Longleftrightarrow \quad F = \tilde{F} \circ J. \qquad \square$$

Lemma G.17 (Row comparison with stochastic emissions)**.** *Suppose $P_a(s, j) \geq cP_a(t, j)$ for every action, destination, and pair of hidden states, where $c > 0$ and $c > 1 - \delta$. Then the model is δ-admissible for all native goals, with arbitrary emissions.*

Proof. We choose $\epsilon > 0$ so that $c(1 - \epsilon) \geq 1 - \delta$. For each goal g and possible initial observation o, we choose a reference state $s_{o,g}$ with $\Omega(o \mid s_{o,g}) > 0$ that maximizes the conditional optimum $V^*_{M,s,o}(g)$ of Section 3. The first action distribution is common across hidden states. Row comparison after that action gives $V^\pi_{M,s,o}(g) \geq cV^\pi_{M,t,o}(g)$ whenever both states can emit o; acceptance before the first transition has the same indicator and obeys this bound because $c \leq 1$, as summing the hypothesis over destinations gives $1 = \sum_j P_a(s, j) \geq c \sum_j P_a(t, j) = c$.

For each goal and o, we choose a policy within the factor $1 - \epsilon$ of the conditional optimum at $s_{o,g}$, using any policy if that optimum is zero. We combine these policies by goal and initial observation. The row comparison and the maximality of $s_{o,g}$ give, for each possible (s, o),

$$V^\pi_{M,s,o}(g) \geq c(1 - \epsilon)V^*_{M,s_{o,g},o}(g) \geq c(1 - \epsilon)V^*_{M,s,o}(g) \geq (1 - \delta)V^*_{M,s,o}(g).$$

Averaging over the initial observation gives

$$V^\pi_{M,s}(g) \geq (1 - \delta) \sum_o \Omega(o \mid s)V^*_{M,s,o}(g) \geq (1 - \delta)V^*_{M,s}(g). \qquad \square$$

With deterministic emissions it suffices to compare states that emit the same observation.

Lemma G.18 (Row comparison with deterministic emissions)**.** *Suppose emissions are deterministic. If, for some $c > 0$, $P_a(s, j) \geq cP_a(t, j)$ for every action a, destination j, and pair of states s, t with the same emitted observation, then the model is δ-admissible whenever $c > 1 - \delta$.*

Proof. We choose $\epsilon > 0$ so that $c(1 - \epsilon) \geq 1 - \delta$, and for each goal g and emitted observation o a reference state $s_{o,g}$ emitting o with the largest optimum $V^*_{M,s}(g)$ among such states. For any policy π and native goal g, conditioning on its private randomness and first action gives

$$V^\pi_{M,s}(g) \geq cV^\pi_{M,t}(g) \quad \text{when } s, t \text{ emit the same observation.}$$

Indeed, acceptance before the first transition agrees at both states; after that transition, the row comparison applies to the same nonnegative continuation values. Summing a transition row also gives $c \leq 1$. For each goal and initial observation o, we choose a policy attaining at least $(1 - \epsilon)V^*_{M,s_{o,g}}(g)$, using any policy when the supremum is zero. These policies combine into one complete goal-conditioned policy because their initial observations distinguish the choices. The row comparison at the chosen policy and the maximality of $s_{o,g}$ give

$$V^\pi_{M,s}(g) \geq cV^\pi_{M,s_{o,g}}(g) \geq c(1 - \epsilon)V^*_{M,s_{o,g}}(g) \geq c(1 - \epsilon)V^*_{M,s}(g) \geq (1 - \delta)V^*_{M,s}(g). \qquad \square$$

The accompanying code includes strictly positive three-state examples with binary emission rank two, Hankel rank three for J, and nontrivial profile-preserving bridges. Their row-comparison coefficients satisfy Lemma G.17 at $\delta = 1/10$; exact matrices and Hankel minors are supplied with the implementation.

Proposition G.19 (Action-dependent resets share exact oracles)**.** *On the communicating model class with joint action image rank one, any two models with the same J share a complete exact statewise oracle.*

Table 2: Competence and access promises. Each row concerns one fixed policy. The sampling guarantees require fresh private randomness.

Target	Competence premise	Access and model scope
Exact information	Optimal for every hidden start and native goal	Complete policy; any admissible subclass
Positive-loss information	Statewise loss $\delta < 1/2$ for every native goal	Complete policy; full finite class
Recovering J	Average loss $\delta < 1/2$ over prescribed words in each queried family; one fixed initial law	Sampled actions at one possible root
Component bit	Statewise competence on the specified loss interval	Nine histories; promised J_0 and four-state cap

Proof. Since $P_a\mathbf{1} = \mathbf{1}$, the common one-dimensional image forces $P_a = \mathbf{1}\mu_a$ for a probability row μ_a. The next observation has law $\nu_a = \mu_a\Omega$, independent of the preceding hidden state and observation. These laws are determined by the conditional probabilities under J, $J(O_1 = o \mid A_0 = a)$.

We choose an optimal policy for each goal in the common controlled process formed by its acceptance monitor and the current observation. Lemma B.2 gives one policy attaining the conditional optimum $v_g(o)$ at every initial observation. Averaging separately in either model therefore gives

$$V^{\pi}_{M,s}(g) = \sum_o \Omega_M(o \mid s) v_g(o) = V^{*}_{M,s}(g). \qquad \square$$

The remaining three-state classification concerns nonminimal realizations and minimal realizations whose joint action image has rank two.

H Interfaces, resources and worked queries

H.1 Calling a policy and observing an environment

At an online revisit to o_*, the agent conditions on its preceding history and the associated hidden-state posterior. Adapting root recovery to this interface requires competence conditional on that history or access to the root policy callable. Corollary D.2 uses competence under one fixed initial law with positive gate probability.

H.2 Competence as a planning interface

Equation (9) concerns one initial law and the prescribed-word distribution of each requested threshold family. It allows lower accuracy on individual words and places no requirement on goals outside those families. A concrete implementation is a planning service whose dynamics model and initial belief are internal: for each finite count goal, it computes the two branch values and returns a maximizing root action. The exact fixed-initial-law planners in Appendix I.8 implement this interface. Exact competence then depends on the planner's model and value computations matching the environment; approximate competence is expressed by (9). The identifier receives neither that model nor the computed values.

H.3 Compatibility and a known state cap

The information result concerns a decoder required to work for every competent policy. If one policy is competent in both M and R, the decoder receives exactly the same input in these two cases, so its target value must agree. A second shared policy transfers this equality from R to N. Thus a chain can force equal guaranteed information even when its endpoints have disjoint policy sets. Proposition 3.5 supplies concrete models with this property.

Table 3: Assumptions and targets of the closest predictive recovery interfaces. The two competence conditions concern different task distributions.

	Native-goal recovery here	Threshold bets in Nayebi (2026, Section 5)
Task supplied	Finite goals on observations and actions; continuation actions are prescribed in the goal	A history-dependent future event compared with a known independent lottery
Response	Sampled environment action at a supplied root observation	A separate report bit that leaves the dynamics unchanged
Competence for extraction	Fixed relative loss $\delta < 1/2$ averaged over prescribed words in each queried family; one initial law	Average regret over evaluation histories and prediction tasks; consistency as regret vanishes
Predictive target	Random-walk law J; all controlled predictions from its stationary initialization	History-conditional predictive probabilities
Linear realization	Classical Hankel rank; a known state cap supplies a finite sufficient menu	Linear PSR operators under an invertible core-matrix assumption
Information converse	Exact admissible subclasses; fixed positive loss on the full finite admissible class	The cited recovery results do not assert equality of complete competent-policy information classes
Principal resources	Goal depth and width, root replies, and oracle computation	Prediction tasks, lottery implementation and report access

For example, set $\delta = 1/4$ and $\alpha_0 = 1/5$ in (17). At a given initial observation, the goal requiring next observation one has (L, R) success probabilities $(1/5, 1/2)$ or $(4/5, 1/2)$ in the two hidden contexts of M. Model N exchanges the actions. Writing θ for the root probability of L, the competence inequalities require

$$\theta \in [1/3, 5/12] \quad \text{in } M, \qquad \theta \in [7/12, 2/3] \quad \text{in } N.$$

These intervals are disjoint. Nevertheless, either model shares a complete competent oracle with the two-state fair reference R_0 in Proposition 3.5. All three models have the same random-walk law. This particular bridge uses only two states.

Bridge size is therefore a property of the realization geometry. Theorem 3.6 allows every finite size; Theorem G.1 restricts all bridges to the declared cap. Its deterministic three-state case preserves the cap because the relevant hulls are points or segments. Theorem G.15 likewise preserves n states for minimal realizations with full joint action image and emission-profile multiplicities at most two. The four-state construction shows how a cap can instead leave two components within one random-walk law.

H.4 Comparison of behavioral identification interfaces

In standard fixed-dynamics inverse reinforcement learning, the transition model is specified and observed decisions constrain an unknown reward (Ng & Russell, 2000). Multiple rewards may induce the same behavior even with unlimited data; the resulting identification class depends on the response information available, such as demonstrations or trajectory comparisons (Skalse et al., 2023). Plan recognition instead infers goals from actions in a supplied domain theory (Ramírez & Geffner, 2009).

Here each goal is specified by the identifier, the environment is unknown, and the oracle can be called on new goal inputs. Equality of complete exact competent-policy sets yields an environment class indexed by J. At positive loss, a single compatible policy need not characterize that entire class: intersections of competent-policy sets generate the compatibility components.

Action equivalence and value queries. Batra (2026, Theorem 3.4) classifies the transition kernels of a fully observed discounted MDP that preserve optimal-action sets for every state–action reward; Theorem 4.2 treats successor-dependent rewards. That reward interface and discounting differ from success of native observation goals. Letcher et al. (2026, Theorems 1–3) instead infer dynamics

from numerical Q-values, policies and known rewards under injectivity or rank conditions. Our oracle supplies actions. The hidden-state setting additionally distinguishes recovery of observable predictions from recovery of labeled transition kernels.

Weiss & Nadler (2015, Section 4.2) characterize equivalent parametric-output HMMs with exactly two aliased states. Their realization geometry concerns equality of observable laws; the positive-loss argument here additionally transfers competence between controlled realizations.

H.5 Predictive channels, latent representations and automata

Grimm et al. (2020) define value-equivalent models through their Bellman updates on selected policy and value-function families. Expanding these families imposes more constraints on the model. The present invariant instead arises from the complete family of native observation goals and the policies meeting their competence promises.

Rosas et al. (2025) study canonical predictive transducers and linear reductions for a supplied environment input/output interface. Baltieri et al. (2026) distinguish environment, agent and joint predictive channels; their joint-model reconstruction determines the environment channel on the supported actions. These distinctions clarify the role of the independent full-support action reference in J. Our recovery problem is to obtain that random-walk law from competent responses, after which the classical predictive representation becomes available.

Several recent works ask what trained latent representations identify. Zhang et al. (2026) obtain representation and controlled-transition identifiability under Gaussian latent-model assumptions, spectral separation and conditional action excitation. Gao et al. (2026) connect prediction and one-step sufficiency through strict mediation by textual latent states. Tan et al. (2026) empirically distinguish physical information available in observations from information decoded from, and used by, multimodal predictive representations. Those claims concern specified model or representation-learning settings. Our target is defined by all competent policies for the supplied goal language, independently of how any one policy was trained or represented.

Observable-law identification for functions of finite Markov chains has classical precedents (Blackwell & Koopmans, 1957; Ito et al., 1992). Ito et al. (1992) characterize equivalence and minimum degrees of freedom through invariant linear subspaces. The controlled four-state result here adds the graph generated by shared approximately competent policies: it proves all-goal admissibility, separates two components with a native goal, and excludes every bridge under the state cap. These conclusions concern behavior guaranteed by competence as well as equality of random-walk laws.

The realization and equivalence steps also have classical precedents. Anderson (1999) studies HMM realization from output-string laws using Hankel matrices and polyhedral cones. Huang et al. (2016) study generic minimal HMM and quasi-HMM realization from finite-length output probabilities. Tzeng (1992) and Kiefer et al. (2012) give algorithms and complexity results for equivalence of supplied probabilistic automata. These works place finite predictive realization and equality testing in an established literature. Our competence-to-cylinder theorem and complete-information converse address access to these representations from a behavioral oracle.

Hidden task memory is another related object. Shehab et al. (2026) infer reward-machine structure and state labels from state trajectories and history-policy information. In our interface the goals are supplied; the unknown object is the environment's predictive dynamics.

H.6 Trajectory data, stopping and continuous observations

Trajectory-based spectral methods estimate cylinder or test probabilities directly from observed sequences (Hsu et al., 2012; Boots et al., 2011). The analysis in Hsu et al. (2012, Section 2.4) uses independent sequences to isolate moment-estimation error; it also discusses using a single long sequence with mixing-dependent effective sample size. Boots et al. (2011) describes a suffix-history construction from a long action–observation sequence when reset is unavailable, with the history reference tied to the data-generating policy. Thus trajectory access itself includes distinct sampling regimes.

Our spectral comparison uses overlapping windows from one stationary-start, uniform-action trajectory per fit. Such observations are dependent. The root interface instead supplies fresh action

draws at selected goals and pays for the competence of the responding policy. Its mixing dependence appears in the goal length through $G_{M,c}$. Both sources can feed Hankel realization, whose stable finite-sample use also depends on rank and conditioning. Access to a competent responder and access to environment transitions are separate resources; Appendix I.7 reports their prediction errors, numerical counts and computation times individually.

At positive loss, Theorem D.6 excludes a uniformly valid, almost surely finite stopping rule on the full finite-goal, fixed-initial-law class without a quantitative mixing promise. The pointwise schedule remains valid without such a promise. With a supplied bound, Theorem 4.3 gives a stopping certificate.

Corollary D.3 uses finite partitions of a known countable alphabet. A continuous alphabet generally has no positive-probability singleton gate, and a countable list of point cylinders no longer determines all required probabilities in the same manner. Extending the sampling result requires a measurable positive-probability gate, a specified determining family of observable tests, and approximation or uniform-convergence control for that family. These are additional interface and regularity conditions.

For continuous actions, a finite grid changes the controlled experiment. Extending the grid's predictions to unqueried actions additionally requires regularity of the transition kernels and competence for the discretized goals. For stochastic emissions, Appendix G.6 extends the three-state characterization to minimal realizations with full joint action image.

H.7 Reference actions and finite resources

H.7.1 Full-support reference actions

Let μ be a distribution of full support on the finite action set. Write $P_\mu = \sum_a \mu(a) P_a$, let η_μ be its stationary row, and define

$$B^\mu_{(o,a)} = \mu(a)\,\mathrm{diag}(\Omega(o \mid \cdot))P_a, \qquad J^\mu_M(w) = \eta_\mu B^\mu_w \mathbf{1}.$$

Communication makes P_μ irreducible. The uniform reference used in the main text is denoted by unif.

Proposition H.1 (Changing the reference action distribution)**.** *For any two communicating models with the common visible interface,*

$$J^\mu_M = J^\mu_N \quad \Longleftrightarrow \quad J^{\mathrm{unif}}_M = J^{\mathrm{unif}}_N.$$

The exact and positive-loss information characterizations therefore hold with J^μ in place of J. The root-action extraction theorems also hold with $W \sim \mu^{\otimes L}$ and the corresponding Poisson constant $G^\mu_{M,c}$, provided the identifier can sample μ.

Proof. We change action probabilities before taking a stationary limit. For a pair word $w = ((o_0, a_0), \ldots, (o_{k-1}, a_{k-1}))$, reweighting gives

$$\mathbb{P}^{\mathrm{unif}}_{M,\eta_\mu}(w) = J^\mu_M(w) \prod_{i=0}^{k-1} \frac{1/m}{\mu(a_i)}.$$

All denominators are positive. Equal J^μ gives equal finite uniform-action trajectory laws from the respective η_μ distributions. We average over time because P_{unif} may be periodic. In particular, the probability of a cylinder c beginning at time t is the same in the two models. Its Cesàro average converges to $J^{\mathrm{unif}}(c)$ because the finite hidden chain P_{unif} is irreducible. Reweighting from uniform actions to μ and taking Cesàro averages under P_μ proves the reverse implication, including periodic chains.

The laws J^μ and J have the same fibers. The information characterizations therefore transfer on their original model classes. For extraction, we draw the prescribed action word from $\mu^{\otimes L}$. Averaging the branch probabilities over these words now gives the independent-μ experiment. The finite window chain has transition weights $P_a(s, s')\Omega(C' \mid s')\mu(a')$ and is irreducible. Its mean-zero Poisson solution supplies the averaging bound, uniformly over the post-marker hidden laws. The threshold proof uses only that bound, giving the same reply margin $\Delta/4$ and hence

$$N \geq \max\left\{2, \frac{256(2-\delta)(G^\mu_{M,c})^2}{\Delta\epsilon^2}\right\}, \qquad K \geq \frac{8}{\Delta^2}\log\frac{B}{\zeta}. \qquad \square$$

The initial-law variant in Corollary D.2 changes the distribution under which competence is evaluated. Here μ changes the reference action experiment; the two choices are distinct.

Full support cannot be omitted in general. Consider two fully observed models on states $\{0,1\}$ with actions L,R. In both, P_{L} has both rows $(1/2,1/2)$; P_{R} is the identity in one model and the swap matrix in the other. Both models communicate and admit exact policies. If $\mu(\mathsf{L})=1$, both reference laws are the same fair IID observation law. At observation 0, the native goal requiring the next observation to be 0 has branch values $(1/2,1)$ in the first model and $(1/2,0)$ in the second. Its exact root action is therefore R in the first and L in the second.

H.7.2 An explicit Poisson constant

Write $n=|S|$. For the reference chain P_μ, define

$$\bar\tau_\mu=\max_{s,t}\mathbb{E}_s\tau_t,\qquad \tau_t=\inf\{j\ge 0:S_j=t\},\qquad \rho_c=\prod_{j=1}^{d}\mu(a_j),$$

where $c=((o_1,a_1),\dots,(o_d,a_d))$. The value of $\bar\tau_\mu$ is zero for a one-state chain. Let D_μ be the directed diameter of the positive transition graph of P_μ, and let κ_μ be its smallest positive transition probability.

Proposition H.2 (Hidden-chain bound for a cylinder)**.** *The mean-zero window-chain constant in Lemma B.1 satisfies*

$$G^{\mu}_{M,c}\le d+\rho_c\bar\tau_\mu.$$

If there is more than one hidden state, then

$$\bar\tau_\mu\le D_\mu\kappa_\mu^{-D_\mu}\le(n-1)\kappa_\mu^{-(n-1)}.$$

These bounds allow periodic chains. If, additionally, for some integer $\ell\ge 1$, number $0<\alpha\le 1$, and probability row λ,

$$P^{\ell}_{\mu}(s,\cdot)\ge\alpha\lambda(\cdot)\quad\text{for every } s,$$

then the alternative bound $G^{\mu}_{M,c}\le d+\rho_c\ell/\alpha$ holds.

Proof. We separate the first d overlapping windows from the hidden chain that follows them. Write Q,ν,f for the window transition matrix, stationary row, and cylinder indicator from Appendix B.2. Let $g(s)=\mathbb{P}^{\mu}_{M,s}(c)$, $\bar g=\eta_\mu g$, and let the row $\Phi(y)$ be the distribution of S_d conditional on the length-d window y. Then $0\le g\le\rho_c\mathbf{1}$ and

$$Q\Phi=\Phi P_\mu,\qquad \nu\Phi=\eta_\mu,\qquad Q^d(f-\bar g\mathbf{1})=\Phi(g-\bar g\mathbf{1}).$$

The final hidden state and action in y determine $\Phi(y)$; after that action, future actions again have law μ.

We choose h with $(I-P_\mu)h=g-\bar g\mathbf{1}$ and $\eta_\mu h=0$. Stopping its Poisson martingale at $\tau_t\wedge r$ and then taking $r\to\infty$ gives

$$h(s)-h(t)=\mathbb{E}_s\sum_{j=0}^{\tau_t-1}(g(S_j)-\bar g).$$

The passage to the limit uses bounded h,g and finite expected hitting times. Finiteness follows directly by following a positive path to t: from every state, the chance to hit t within D_μ steps is at least $\kappa_\mu^{D_\mu}$. Repeating this bound at block boundaries gives $\mathbb{E}_s\tau_t\le D_\mu\kappa_\mu^{-D_\mu}$; the one-state case is immediate. Since $\bar g$ lies between the minimum and maximum of g, $\max h-\min h\le\bar\tau_\mu(\max g-\min g)$. Mean-zero normalization gives $\|h\|_\infty\le\rho_c\bar\tau_\mu$, since every weight of η_μ is positive and its weighted average of h is zero.

The mean-zero window solution is

$$u=\sum_{j=0}^{d-1}Q^j(f-\bar g\mathbf{1})+\Phi h.$$

Indeed, applying $I - Q$ telescopes to $f - \bar{g}\mathbf{1}$, and $\nu u = 0$ by the three identities above. Each term in the sum has norm at most one, while Φ is stochastic, so $\|u\|_\infty \le d + \rho_c \bar{\tau}_\mu$.

Under the additional minorization, we couple two copies at each block of ℓ steps using their common $\alpha\lambda$ component. This gives $\sup_s \|P_\mu^j(s,\cdot) - \eta_\mu\|_{\mathrm{TV}} \le (1-\alpha)^{\lfloor j/\ell \rfloor}$. The convergent series $h = \sum_{j\ge 0} P_\mu^j(g - \bar{g}\mathbf{1})$ is the same mean-zero solution. Bounding each term by ρ_c times this total variation distance yields

$$\|u\|_\infty \le d + \|h\|_\infty \le d + \rho_c \sum_{j\ge 0} (1-\alpha)^{\lfloor j/\ell \rfloor} = d + \rho_c \ell/\alpha. \qquad \square$$

For uniform actions, $\rho_c = m^{-d}$. A known bound on $\bar{\tau}_{\mathrm{unif}}$, or on n and the positive transition probabilities, makes the sufficient goal length in Theorem 4.3 explicit. Communication alone supplies no uniform numerical value for these parameters. The Poisson stopping identity is standard; a general treatment is given by Glynn & Infanger (2023, Section 4, Theorem 2).

H.7.3 Count-monitor computation

Let k_c be the cylinder's observation-cell count, $L = N + d - 1$, and $k = \lceil Nq \rceil$, with $0 < q \le 1$. A prefix monitor records the longest proper prefix of c that is a suffix of the pairs already read. Write $T(r,x)$ for its next prefix state and $I(r,x) \in \{0,1\}$ for the indicator that the new pair completes a copy of c. The prefix states are $r \in \{0,\dots,d-1\}$, including after overlapping matches.

Proposition H.3 (Cost of evaluating the prescribed count event)**.** *Given its prefix transition table, the event in $g(c,N,q,w)$ is evaluated by a monitor with $O(Ld(k+1))$ states. A model-aware evaluator computes the event probability from all n possible hidden starts in $O(Ln^2 k_c d(k+1))$ arithmetic operations and $O(nd(k+1))$ working numbers, in addition to the model, prescribed word, and prefix table. The two root-marker branch values require a further $O(n^2)$ operations.*

Proof. We store the time position, prefix state, root marker, and match count capped at k. A departure from w rejects the goal. After the L prescribed pairs, the positive branch accepts at count k, and the negative branch accepts below k. The prefix update retains overlaps, so the count equals the number of successful length-d windows. These variables give the monitor-state bound.

For the probabilities, we use a backward array $V_j(s,r,b)$ for the chance of ending at capped count k, given the next hidden state s, prefix r, and count b after the first j pairs. We set $V_L(s,r,b) = \mathbf{1}\{b = k\}$. With $a = w_{j+1}$, the recursion is

$$V_j(s,r,b) = \sum_C \Omega(C \mid s) \sum_t P_a(s,t) V_{j+1}\big(t, T(r,(C,a)), \min\{k, b + I(r,(C,a))\}\big).$$

The final transition has no effect because V_L is independent of the hidden state. We keep two time slices. Every slice contains $nd(k+1)$ entries, and every entry sums at most $k_c n$ terms. All starting values $v(s) = V_0(s,0,0)$ use the claimed time and memory. Applying the two root transitions gives

$$A_s = (P_{a_+} v)(s), \qquad B_s = (P_{a_-}(\mathbf{1} - v))(s). \qquad \square$$

The monitor evaluates exactly the same accepting paths as the expanded native disjunction, so its goal values and competence margins are unchanged. The simulator stores the prescribed word, prefix state and capped count; it never materializes the exponentially many disjuncts. The arithmetic bounds are polynomial in the numerical horizon and threshold, conditional on known model parameters. Horizon dependence, arbitrary-agent planning cost and exact-arithmetic bit complexity remain separate resources. The binary implementation uses $k_c \le 2$. Fully observed models use forward recurrences, aliased models use the backward array, and deterministic models need one path per compatible initial state and marker.

H.7.4 Contaminated root replies

Consider a qualified threshold comparison from Lemma 4.2. Let X_i indicate the clean root reply a_+ and $\widetilde{X}_i$ its observed value. Conditional on all past queries, each fresh word and clean reply satisfy the lemma. A contamination flag has conditional probability at most ϱ and is independent of that

fresh word and clean reply; a flagged reply may be replaced by any bit chosen from current and past query information. Fresh clean replies are sampled conditionally on the current goal. The corruption changes replies, not prescribed words.

Proposition H.4 (Recovery under independent response contamination)**.** *Suppose the contamination model above holds and the known upper bound satisfies $\varrho < \Delta/(2+\Delta)$. Write*

$$\Delta_\varrho = (1-\varrho)\Delta - 2\varrho > 0.$$

Above a qualified threshold, $\mathbb{E}[\widetilde{X}_i \mid \text{past}] \geq 1/2 + \Delta_\varrho/4$; below it, this expectation is at most $1/2 - \Delta_\varrho/4$. The majority test with K replies has error at most $\exp(-K\Delta_\varrho^2/8)$. The finite recovery theorem consequently holds with its original N, B and

$$K = \left\lceil \frac{8}{\Delta_\varrho^2} \log \frac{B}{\zeta} \right\rceil.$$

Proof. We bound the adversary by setting every corrupted positive-side reply to zero, and every corrupted negative-side reply to one. Conditional independence of the flag preserves the clean mean on unflagged draws. The two bounds are therefore

$$(1-\varrho)(1/2+\Delta/4) = 1/2 + \Delta_\varrho/4, \qquad (1-\varrho)(1/2-\Delta/4) + \varrho = 1/2 - \Delta_\varrho/4.$$

For a bounded variable, the conditional exponential-moment bound used in Lemma 4.2 applies at each draw. Iterating these conditional bounds gives the majority error exponent, even when replacement bits depend on earlier queries. A union bound over the B comparisons, followed by the bisection argument, gives

$$\mathbb{P}(|\widehat{J}(c) - J_M(c)| > \epsilon) \leq B\exp(-K\Delta_\varrho^2/8) \leq \zeta. \qquad \square$$

The threshold on ϱ is exact for preservation of the known clean mean interval under this contamination model; it is not a lower bound for every native extraction algorithm. Without independence of the flag, replacing at most ϱK replies in a batch changes its empirical mean by at most ϱ. The same argument then gives error at most $\exp(-2K(\Delta/4-\varrho)^2)$ when $\varrho < \Delta/4$. Arbitrary competent tie-breaking already lies within Lemma 4.2; contamination adds possible failures of the competence promise. No averaging guarantee follows for unrestricted dependence among clean replies that destroys their conditional mean bounds.

H.8 A NATIVE QUERY AND ITS IMPLEMENTATION

H.8.1 A THREE-WINDOW QUERY

Take observations $0, 1$, actions a, b, and supplied root observation o_*. For cylinder $c = (1, \mathsf{a})$, window count $N = 3$, threshold $q = 2/3$ and continuation word aaa, partition the observation triples as

$$\mathcal{Z}_+ = \{011, 101, 110, 111\}, \qquad \mathcal{Z}_- = \{000, 001, 010, 100\}.$$

For each triple $z = z_1z_2z_3$, its disjunct is

$$\text{current}\{(o_*, a_z)\}; \quad \text{next}\{(z_1, \mathsf{a})\}; \quad \text{next}\{(z_2, \mathsf{a})\}; \quad \text{next}\{(z_3, \mathsf{a})\}, \qquad a_z = \begin{cases} \mathsf{a}, & z \in \mathcal{Z}_+, \\ \mathsf{b}, & z \in \mathcal{Z}_-. \end{cases}$$

Their disjunction has depth four and width eight. Root action a selects the event of at least two occurrences of $(1, \mathsf{a})$, while b selects its complement. Every branch prescribes its continuation actions. For example, replies $\mathsf{a}, \mathsf{a}, \mathsf{b}, \mathsf{a}, \mathsf{b}$ give an event-side majority, determining the half-interval retained by bisection.

This is one explicit goal in the query family. The extraction theorem samples continuation words uniformly and uses the length and reply bounds in Theorem 4.3. The two markers can change the hidden state before the counted segment; the uniform-over-starts averaging bound handles this effect.

Table 4: Resources for one cylinder at the finite-accuracy setting. Here $L = N + d - 1$, $k = \lceil Nq \rceil$ and $n = |S|$. The last two rows are costs for a simulator with known model parameters.

Resource	Sufficient cost
Root replies	BK
Native goal depth / width	$N + d$ / k_c^{N+d-1}
Count-monitor states	$O(Ld(k+1))$
Simulator arithmetic per event	$O(Ln^2 k_c d(k+1))$
Simulator working numbers	$O(nd(k+1))$

H.8.2 Count-monitor evaluation

For a general cylinder of length d, the monitor stores time, the matched cylinder prefix, root marker and occurrence count capped at $k = \lceil Nq \rceil$. Its acceptance rule is:

1. Check the root observation and store the selected marker.
2. At each continuation position, reject an action that differs from the prescribed word.
3. Update the matched prefix and increment the capped count when a cylinder ends; the prefix state retains overlaps.
4. After $L = N + d - 1$ positions, accept the event marker at count k and the complement marker below k.

For the example above, each occurrence of $(1, \mathsf{a})$ increments the count. Proposition H.3 gives the general prefix automaton and derives the computational bounds in Table 4.

H.8.3 Representative admissible classes

Fully observed finite models admit policies optimal on the product of the visible state and finite goal monitor. The three-state example in Section 2 also admits exact statewise policies, while retaining predictive information in history. With deterministic observations and stochastic transitions, Lemma G.18 gives a sufficient condition: transition rows within an observation group dominate one another by a factor c satisfying $c > 1 - \delta$. The stochastic forecasting models belong to this family. Table 2 distinguishes these statewise requirements from the fixed-initial-law premise for root-action recovery.

I Experiments

I.1 Fully observed and copied models

The observation and action sets are $\{0, 1\}$. The IID model is fully observed with identical transition rows $(0.65, 0.35)$ for both actions. The controlled model is fully observed with

$$P_0 = \begin{pmatrix} 3/4 & 1/4 \\ 1/3 & 2/3 \end{pmatrix}, \qquad P_1 = \begin{pmatrix} 1/2 & 1/2 \\ 1/5 & 4/5 \end{pmatrix}. \tag{76}$$

The four-state realization uses (14) with

$$W_0 = \begin{pmatrix} 4/5 & 1/5 \\ 3/10 & 7/10 \end{pmatrix}, \qquad W_1 = \begin{pmatrix} 2/5 & 3/5 \\ 9/10 & 1/10 \end{pmatrix}.$$

The two copies of each visible state have the same controlled observation law, so no hidden-copy information is needed for statewise optimality. Each fully observed quotient combined with the finite native monitor supplies an exact all-goal policy.

For the queried goal, the oracle computes the two branch probabilities by dynamic programming over the hidden state, the cylinder-prefix matching state, and the number of cylinder occurrences. The count is capped at the threshold, and overlapping occurrences are retained. An a_+ root followed by

the prescribed word accepts the upper tail; an a_- root accepts its complement. At loss zero the oracle selects a_+ when its branch value is at least the other. At positive loss, if $A_w > B_w$, it uses

$$\theta_w = \max\left\{0, \frac{(1-\delta)A_w - B_w}{A_w - B_w}\right\};$$

otherwise it uses zero a_+ probability. It then follows the written continuation. This minimizes the root probability of a_+ subject to the policy-average competence inequality. The rule is a function of the queried goal and model, independent of the identifier's current estimate. On initial observations other than the gate, the goal's value is zero. The model's lumpability makes the queried conditional values the same at every hidden state with that initial observation.

All root queries use observation zero. For each reply the identifier draws a new independent uniform action word and the oracle returns a fresh action sample. The identifier receives no transition or emission matrices, branch values, stationary probabilities, or realized environment trajectories. The evaluation routine has the model parameters to score predictions.

I.2 Predictive realization

Let $\widehat{\mathsf{H}} = U\Lambda V^\top$ be the SVD of the estimated Hankel menu. We retain the r singular values exceeding $4N^{-3/4}$, at least one, with the matching blocks U_r, Λ_r, V_r. The realization is

$$\widehat{\alpha} = \widehat{\mathsf{H}}(\varnothing, :)V_r\Lambda_r^{-1/2}, \qquad \widehat{\beta} = \Lambda_r^{-1/2}U_r^\top\widehat{\mathsf{H}}(:, \varnothing), \qquad \widehat{T}_x = \Lambda_r^{-1/2}U_r^\top\widehat{\mathsf{H}}_x V_r\Lambda_r^{-1/2}.$$

The true rank is used for scoring only. The complete menu below uses prefixes and suffixes through length two. The general consistency result uses the state-cap menu and determinant rule in Appendix F.

I.3 A common menu including the three-state model

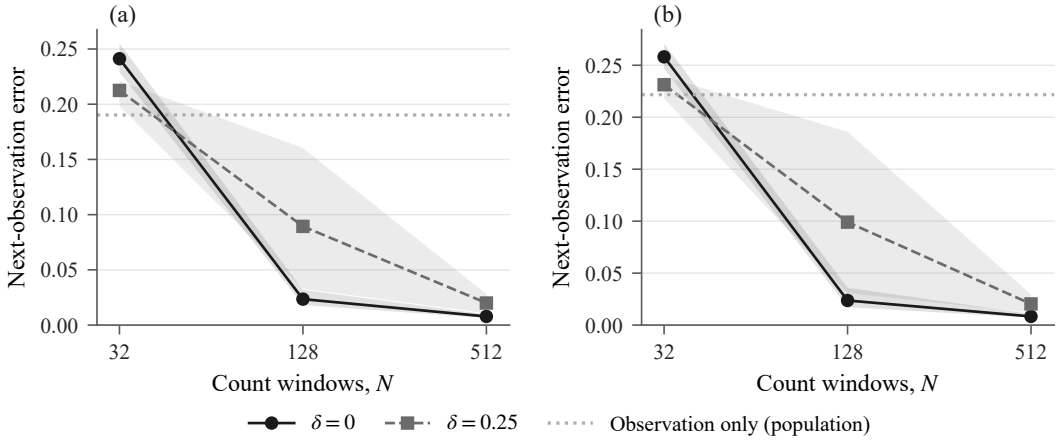

Figure 4: **Root actions recover information available from history.** Three-state prediction error on (a) all 240 held-out histories and (b) the 206 histories ending in the aliased observation x whose hidden state is determined by the history. Points average five seeds; bands show their range. The dotted predictor uses the exact conditional probability under J given only the current observation and pending action. Each run uses 92,752 root replies.

The common-menu experiment uses the IID and controlled kernels in (76), the three-state model of Section 2, and the relabeling of Proposition B.4. The latter two models are compared with coupled seeds, never pooled as independent repetitions. Their observations x, y and actions a, b are encoded as zero and one.

The three-state model has stationary hidden distribution $(4, 2, 1)/7$ under uniform actions. The Hankel block on all prefixes and suffixes of pair length at most one has rank two. With $\mathcal{U} =$

Table 5: Prediction error at $N = 512$, with correct-rank counts out of five in parentheses. The three-state relabeling uses coupled queries and replies; its row is not an independent replication.

Model	Hidden states	Predictive rank	$\delta = 0$	$\delta = 0.25$
IID	2	1	0.00191 (5)	0.02174 (5)
Controlled	2	2	0.00587 (5)	0.05045 (5)
Three-state	3	3	0.00782 (5)	0.01999 (5)
Three-state, relabeled	3	3	0.00782 (5)	0.01999 (5)

$\{\varnothing, (\mathsf{x},\mathsf{a}), (\mathsf{x},\mathsf{a})(\mathsf{x},\mathsf{a})\}$, the larger menu contains

$$\mathsf{H}(\mathcal{U},\mathcal{U}) = \begin{pmatrix} 1 & 3/7 & 1/7 \\ 3/7 & 1/7 & 0 \\ 1/7 & 0 & 0 \end{pmatrix}, \qquad \det \mathsf{H}(\mathcal{U},\mathcal{U}) = -1/343.$$

The three-state factorization bounds the full rank by three, so this minor establishes predictive rank three. The 21-word complete menu through length two supplies this block and all one-letter insertions using cylinders through length five. These finite rational identities are retained as an exact matrix certificate with the experimental artifacts.

Statewise oracle. For each prescribed word w and hidden state s that emits the root observation, let $A_{w,s}, B_{w,s}$ be the upper- and lower-tail branch values after the two root markers. The common feasible root probabilities form

$$\Theta_w = \bigcap_{s:\Omega(o_*|s)>0} \{\theta \in [0,1] : \theta A_{w,s} + (1-\theta)B_{w,s} \geq (1-\delta)\max\{A_{w,s}, B_{w,s}\}\}. \tag{77}$$

The oracle chooses $\max \Theta_w$ at $\delta = 0$ and $\min \Theta_w$ at positive loss, followed by the prescribed continuation. On initial observations outside the gate the goal has value zero. The interval is computed separately for every goal from all compatible hidden starts. An average over these starts would impose a different competence condition.

For the fully observed models, branch values use the count-monitor dynamic program. For the deterministic three-state models, trajectory evaluation gives both values separately at every compatible root. Both implementations retain overlapping cylinder occurrences.

Configuration and evaluation. The same N, loss, seeds, bisection depth, replies per threshold, menu, and SVD rule are used for all four models. There are $(4 + 4^2 + \cdots + 4^5)/2 = 682$ queried cylinders. Every bisection compares on $[0, 2^{-d}]$ for a length-d cylinder, using a new independent uniform action word for each reply. Eight bisections and 17 replies give 92,752 replies per run, with maximum depth $N + 5$. The estimator's inputs are exclusively the sampled root actions.

All words of lengths six and seven are evaluated, including those impossible under the true model. Conditional evaluation uses every pair history of length five or six with positive true probability. There are 240 such histories for the three-state law, of which 206 end in observation x with a uniquely determined current hidden state. The history sets for each fully observed law each contain 5,120 words. Histories are weighted equally within each set. The conditional estimate is $2\widehat{J}(h(\mathsf{y},\mathsf{a}))/\widehat{J}(h)$; the final action in h is the pending intervention. A nonpositive, nonfinite, or at-most-10^{-14} denominator is undefined. The primary score assigns it loss one. All other ratios retain their absolute errors without projection onto $[0, 1]$.

The observation-only predictor uses

$$p_{\mathrm{obs}}(o, a') = \frac{2J((o,a')(\mathsf{y},\mathsf{a}))}{J(o,a')}.$$

It knows the current observation and pending action, but not the earlier history. Its population form is computed only by the evaluator. A second baseline replaces J by the identifier's estimates on the short cylinders, without additional replies. For the three-state model at $N = 512$, this estimated baseline has mean MAE 0.19307 at exact competence and 0.19062 at loss 0.25; the population baseline has MAE 0.19028.

Results. Figure 4 shows the three-state prediction errors as goal length increases. Table 5 reports all model/loss settings at $N = 512$. The population SVD control recovers predictive ranks one, two and three. Thirty coupled relabeling pairs give identical query/reply transcripts and cylinder estimates. The 120-run grid uses 11,130,240 root replies and 596.93 CPU-host seconds.

I.4 Stochastic models and reply resources

I.4.1 Models and fixed protocol

The stochastic models below satisfy the statewise competence promise. Their finite-context predictions are evaluated in Appendix I.5. Root queries use the common goal syntax and independently sampled actions.

Our stochastic examples have $n \in \{3, 4\}$ states numbered $0, \dots, n-1$, two actions $a \in \{0, 1\}$, and observations zero at states $0, \dots, n-2$ and one at state $n-1$. For the zero-emitting states, the transition rows are

$$P_a(s, j) = \frac{1}{n} + \frac{\mathbf{1}\{j = (s+1+a) \bmod n\} - \mathbf{1}\{j = (s+a) \bmod n\}}{10n}.$$

The remaining row assigns probability $19/20$ to state a and $1/[20(n-1)]$ to each other state. Every transition entry is positive. The same-observation row ratios are at least $9/11$, so Lemma G.18 gives admissibility at $\delta = 0.35$. The histories $((0,0),(0,0))$ and $((1,0),(0,0))$ have the same terminal observation and pending action but different next-observation predictions. The finite rational certificates also show Hankel ranks three and four on the common prefix/suffix menu of length at most two.

I.4.2 Reply-resource sensitivity

The dependence on competence loss is visible in the three-state memory model. At $N = 512$ and $\delta = 0.4$, increasing replies per threshold from 17 to 107 reduces mean prediction error from 0.07605 to 0.02936 across three paired repetitions. The increase follows the relative Δ^{-2} scaling in Theorem 4.3, with total reply counts 92,752 and 583,792. At the larger budget all three repetitions select rank three.

Reply contamination acts separately from goal length. At $N = 512$, nominal loss 0.25 and 17 replies per threshold, independent corruption indicators with probabilities $0, 0.05, 0.15, 0.30$ give mean errors $0.01887, 0.02412, 0.07257, 0.37480$, respectively. Marked replies are replaced against the stationary threshold side. Proposition H.4 states the applicable margin condition for this response model.

The simulator evaluates count monitors and includes their dynamic-programming cost in runtime. Maximum native goal depths at $N = 128, 512$ are 133 and 517, with binary disjunction widths $2^{132}, 2^{516}$. Table 4 gives the operation and memory bounds. Appendix I.6 compares behavioral and trajectory acquisition for the final finite-context predictor.

I.5 Behavioral forecasting under stochastic aliasing

We estimate next-observation probabilities from two-pair observable contexts on three stochastic hidden models.

I.5.1 Models and policy controls

The first two models are the three- and four-state aliased kernels in Appendix I.4. Observations are deterministic; the hidden transitions are stochastic. For a third four-state model, the model generator draws integers e uniformly from $85, \dots, 100$ and b from $900, \dots, 960$. Set $\epsilon = e/1000$ in the aliased rows and give the singleton-observation row action-dependent concentration $\beta = b/1000$, dividing its residual mass equally among the other three states. The sampled parameters are $\epsilon = 47/500$ and $\beta = 479/500$. All entries are positive, and the same-observation row ratio satisfies $c \geq 9/11 > 1 - 0.35$. Lemma G.18 supplies all-goal admissibility. The same three models are used in each repetition.

The simulator computes the root-action interval satisfying every compatible hidden-start constraint in (77). Least-left and greatest-left choose its lower and upper endpoints for physical action zero.

Event-adverse selects the endpoint minimizing the upper-tail event, including when the physical markers are exchanged. The cylinder dictionary and predictor settings are fixed before evaluation.

I.5.2 NATIVE QUERIES AND CALIBRATION

For a length-d pair cylinder whose terminal action is zero, the queried counting event unites both terminal-action variants. Its stationary frequency is $2J(c)$ because the last uniform action does not affect its observation. The goal still prescribes the full continuation action word. The two root orientations assign the upper-tail event and its complement to opposite physical root actions, each of which can change the hidden state.

We search integer occurrence cutoffs with $N = 2048$ windows, 257 independent uniform action words per cylinder, and five root samples per word, orientation and cutoff. If k is the largest accepted cutoff, the union-event crossing estimate is k/N and the cylinder estimate is $k/(2N)$. The menu has $(4 + 4^2 + 4^3)/2 = 42$ representative cylinders, with their other terminal-action variants copied. It includes the eight length-two cylinders needed for the one-pair predictor. The public search range uses $2J(c) \le 2^{1-d}$.

For each orientation separately, write u_c for the measured cylinder mass. Within a fixed action word, using terminal action zero, choose the common scalar λ such that

$$\sum_{o_1,\dots,o_d} \left(\max\{0, \sqrt{u_{(o_1,a_1)\cdots(o_d,0)}} + \lambda\}\right)^2 = 2^{-d}.$$

The squared terms define calibrated masses. We copy the terminal-action counterparts and average the two calibrated orientations. The square-root shift is motivated by finite-count quantile bias and uses the public action marginal and measured cylinder masses.

For a suffix h of at most two observed pairs, the predictor is

$$\widehat{p}(1 \mid h) = \frac{\widehat{J}(h(1,0))}{\widehat{J}(h(0,0)) + \widehat{J}(h(1,0))}.$$

The length-two suffix is used only if both uncalibrated, orientation-averaged successor masses multiplied by $2N$ are at least four. Otherwise the predictor uses the last pair. Zero denominators trigger further backoff. Probability normalization is part of this estimator. Context order and count threshold are fixed across models, policies and repetitions.

Each repetition uses separate random streams for action-word generation and root sampling, paired across models and policies.

I.5.3 EVALUATION AND PAIRED REPETITIONS

All positive-probability histories of five and six pairs receive equal weight, giving 5,120 targets per cell. Equal weighting scores every history string once; under J these strings differ in probability by orders of magnitude, so the average emphasizes rare contexts, where backoff is most frequent. A J-weighted average is not reported. The behavioral observation baseline uses the same calibration and normalization with context order one; its eight-cylinder sub-menu has a lower acquisition cost. Two population baselines are evaluator-only: the conditional probability under J given the current observation and pending action, and the per-current-pair median of the target probabilities under this evaluation distribution. The latter minimizes the reported absolute error among all current-pair predictors.

Figure 3 reports all model–policy comparisons. All 27 fits produce defined predictions in $[0, 1]$. The two-pair predictor has lower MAE than the population current-pair baseline in 27 of 27 fits and than the best current-pair predictor for absolute error in 26 of 27.

I.5.4 COSTS AND UNCERTAINTY

Each seed draws 42×257 action words, reused across models and policy controls: globally 32,382 word draws in 126 batches, evaluated in 97,146 model–word instances. Each two-pair fit uses 1,002,300 root replies, giving 27,062,100 replies across the batch; its one-pair sub-menu uses 205,600 replies per fit.

The complete batch takes 2176.55 seconds of host wall time and 4131.50 process CPU seconds, with at most four CPU threads. Count-law computation takes 1595.75 wall seconds, amortized across the three policy controls for each model and seed. Maximum measured process RSS is 512.6 MiB. The longest forced word has length 2050 and the native goal depth is 2051. For the binary observation alphabet, direct disjunction expansion has width 2^{2050}. The simulator evaluates acceptance with the count monitor.

Figure 3 reports empirical ranges across the three paired repetitions, with action words shared across cutoffs and orientations.

I.6 Components of the finite-context predictor

We compare the predictor's components on the same 27 fitted cylinder sets and 5,120 history targets per set as in Appendix I.5. The comparison includes every model, seed and planner control. It crosses three cylinder transformations, three marker choices, context orders one and two, and count thresholds zero and four.

Ordinary normalization multiplies every measured cylinder in a fixed action-word group by $2^{-d}/\sum_c u_c$, leaving an all-zero group at zero. Both calibrations transform each orientation separately before averaging when both markers are used. Reliability uses the selected marker's uncalibrated counts or their uncalibrated average for two markers. Threshold zero retains zero-denominator backoff; the count threshold does not affect order one.

Table 6: Component settings on the same acquired cylinders. MAE averages the 27 fits. Columns headed order 2 use count thresholds zero and four, respectively; the last column reports their backoff percentages. The order-one results are identical at both thresholds. Every setting has zero undefined and zero out-of-range predictions.

Transformation	Marker	Order 1	Order 2, 0	Order 2, 4	Backoff 0/4 (%)
None	Zero	.006647	.010303	.006719	2.8 / 47.2
None	One	.006653	.010201	.006578	4.2 / 47.2
None	Both	.004961	.007394	.002802	1.2 / 50.0
Ordinary	Zero	.006647	.010303	.006719	2.8 / 47.2
Ordinary	One	.006653	.010201	.006578	4.2 / 47.2
Ordinary	Both	.004978	.007396	.002846	1.2 / 50.0
Square root	Zero	.004700	.024497	.002189	0.0 / 47.2
Square root	One	.004593	.030171	.002130	0.0 / 47.2
Square root	Both	.004348	.018982	.001256	0.0 / 50.0

With two markers and four-count backoff, square-root calibration reduces MAE from 0.002802 to 0.001256, compared with 0.002846 for ordinary normalization. It also improves on either calibrated single-marker predictor. The count rule is essential to this gain: its removal raises calibrated error to 0.018982. The complete predictor uses a shorter context for half of the evaluated histories.

For cylinder accuracy, we average $2^{|c|}|\widehat{J}(c)-J(c)|$ over the 42 queried cylinder representatives and all fits, using J only for evaluation. The two-marker errors are 0.003736 before transformation, 0.001046 after ordinary normalization, and 0.000689 after square-root calibration. Square-root calibration can assign positive mass to a measured zero. The reliability test therefore uses uncalibrated counts when selecting the context length.

The same predictor from trajectories. Applying the two-pair successor ratio and four-count backoff directly to observed cylinder frequencies gives MAE 0.00130–0.00157 at 10^6 transitions, using all three seeds for each of the three models. These context estimates complement the matched-count Hankel/PSR comparison in Appendix I.7.

I.7 Trajectory spectral learning at matched numerical counts

We compare the calibrated two-pair root predictor with a trajectory-based Hankel/PSR learner using 1,002,300 root replies or environment transitions. Each transition exposes an observation–action pair.

The trajectory learner receives no hidden states, transition matrices, stationary probabilities or test targets.

For each model we generate three stationary-start trajectories under independent uniform actions. Overlapping frequencies estimate the 682 representative cylinders through length five, pooling terminal-action counterparts. Prefixes and suffixes through length two give a 21×21 Hankel matrix. Balanced truncated SVD produces the initial vector, terminal vector and four pair-symbol operators of a linear predictive representation. The root predictor uses its 42 measured cylinders through length three; the trajectory observations supply the longer menu needed for spectral system learning. Finite defined predictions are projected onto $[0, 1]$. A nonpositive or numerically undefined denominator triggers the observed current-pair next-observation frequency, with the empirical marginal and then $1/2$ as fallbacks for missing data.

Truncation uses observed-data validation. We estimate the candidate representations on the first 80% of a trajectory and select a rank from one to the declared state cap by squared prediction error on the remaining observations, using histories of five and six pairs and the same projection and frequency fallback. Candidate fits and their fallback frequencies use only the training segment. We then refit the selected rank and fallback using the full trajectory. The public state cap is three for the three-state model and four for both four-state models. It supplies a candidate set without supplying a true predictive rank. No evaluation probability selects a rank or threshold. Both methods are evaluated on all positive histories of five and six pairs. Every test prediction is finite and lies in $[0, 1]$.

Table 7: Final forecasting methods at 1,002,300 root replies or transitions. Root MAE averages all three policies and seeds. PSR MAE and sample SD use all three trajectory seeds; the last column gives their validation-selected ranks. Root replies and trajectory transitions remain distinct access units.

Model	Root MAE	Trajectory PSR MAE (SD)	Selected ranks
Three-state	.00135	.00377 (.00075)	3, 3, 3
Four-state	.00119	.00412 (.00153)	3, 2, 2
Four-state draw	.00123	.00732 (.00538)	4, 2, 3

The root predictor has lower MAE on the three tested models. Mean trajectory sampling-and-fitting times, including rank selection, are 0.704, 0.405 and 0.412 seconds. The corresponding root times are 58.24, 90.02 and 89.00 seconds, including the synthetic oracle's count-law computation. Root goals have depth 2051 and direct expansion width 2^{2050}; the simulator evaluates their count monitors.

I.8 Larger hidden-state models with a fixed-initial-law planner

We consider six controlled models: two draws each with 6, 8 and 12 hidden states, two actions and two observations. Each observation identifies half the hidden states. For each observation–action group, a positive reference transition row b is perturbed as

$$P_a(s, j) = b_j \left[1 + 0.095 \left(z_{sj} - \textstyle\sum_k b_k z_{sk}\right)/2\right], \qquad |z_{sj}| \le 1.$$

The two groups have separate reference rows; within a group, the centered perturbations retain stochasticity and positivity. Hidden-state counts describe the model family and are not supplied to the predictor.

The planner starts from the stationary distribution η_M under uniform actions. For a query with initial observation zero, it chooses the action maximizing the branch value averaged over $\mathbb{P}_{\eta_M}(S_0 = s \mid O_0 = 0)$. A count monitor evaluates these values. Competence is exact for this fixed initial law. The identifier receives only root actions, the binary alphabets and the sampling configuration.

Each model uses three independent word seeds, $N = 1024$, 129 words per cylinder and one root reply per word and cutoff. The exact planner returns the same action on repeated copies of a query, making a single reply sufficient at each fixed word. The 42-cylinder menu, both marker orientations, square-root calibration and four-count backoff are fixed. All fitted cylinders are saved before evaluation on the 5,120 histories of five and six pairs. The one-pair comparison uses the same measured cylinders and no additional queries.

Table 8: Forecasting with the fixed-initial-law planner. Each of the six models includes all three word seeds; entries give means and sample SD. One-pair uses the measured cylinders; best-pair uses evaluator-only medians within each current-pair group. Time is mean wall seconds per fit on CPU, including count-law computation.

States / draw	Two-pair MAE (SD)	One-pair MAE	Best-pair MAE	Seconds
6 / 1	.00316 (.00055)	.00527	.00527	68.3
6 / 2	.00455 (.00059)	.00793	.00793	66.9
8 / 1	.00376 (.00024)	.00624	.00624	118.7
8 / 2	.00412 (.00047)	.00632	.00632	118.2
12 / 1	.00362 (.00042)	.00581	.00571	268.0
12 / 2	.00251 (.00040)	.00153	.00129	194.3

At each state count, averaging both draws and all seeds gives lower two-pair MAE than either current-pair comparison. Across all 18 fits, mean MAE is 0.00362 versus 0.00546 for the best current-pair predictor, a 33.7% reduction. Each fit uses 89,784 root replies and 5,418 word draws, with goal depth at most 1027 and native width 2^{1026}. The batch takes 2,502.95 wall seconds with four CPU workers. The second 12-state draw is the exception: its two-pair MAE .00251 exceeds both current-pair comparisons (.00153 and .00129). It has the smallest current-pair error of the six models, so the least history dependence to recover, and the length-two estimates at this budget do not improve on it.