Goal-Optimal Agents Necessarily Learn Predictive World Models in POMDPs with episodic resets

Keshav Goyal [*1] **Suraj Yadav** [*2]

Abstract

We extend the recent "General agents contain world models" result of Richens et al. (2025) from fully observable MDPs to partially observable settings under the assumption of episodic resets. Using resets to regenerate histories and enable repeated probing, we show that any agent optimal on a sufficiently rich class of goals must implicitly encode a predictive model of its environment's observation dynamics. This extraction is robust to imperfect resets: accepting approximate neighborhoods of a target belief state rather than regenerating histories exactly introduces only a small, bounded error while yielding an exponential reduction in sample complexity.

1. Introduction

The interplay between perception, representation, and action is central to artificial intelligence. A key question in this field is whether world models (explicit internal representations of environmental dynamics) are necessary for flexible, goal-directed behaviour, or if model-free reinforcement learning can ultimately suffice.

A longstanding motivation for goal-conditioned reinforcement learning (Liu et al., 2022) is the hope of bypassing explicit model learning, as estimating full transition probabilities can be computationally challenging. Rather than estimating transition parameters and subsequently planning, one hopes to train an agent that is directly competent at achieving a rich family of goals—such as reachability, sequencing, or success-event predicates. This promise is particularly attractive under conditions of sparse rewards, non-stationary downstream tasks, or when success detectors are easier to specify than full dynamics (Florensa et al., 2018).

Recent work of (Richens et al., 2025) formally addressed this by proving that world models are a necessary ingredient for agents capable of generalizing to multi-step tasks in fully observable Markov Decision Processes (MDPs). Their core result demonstrates that any agent satisfying a regret bound for a diverse set of goals must have learned an accurate predictive model of its environment. By constructing a small set of composite "switch" goals (see Section 4.1 for more details) , they showed that an agent's preferences uniquely determine the transition kernel– the probability of reaching the state s' once the agent takes action a at the state s– up to a small estimation error. This implies that the hope of avoiding dynamics altogether fails in MDPs; optimality on even a tiny set of goals forces implicit knowledge of the transition kernel.

A natural and critical extension to the MDP result is whether this impossibility persists under partial observability. POMDPs provide the necessary framework for modeling realistic scenarios where perfect state information is unavailable. In such environments, agents face the challenge of *perceptual aliasing*, where distinct underlying states produce identical observations—such as a robot encountering visually identical corridors or a physician observing shared symptoms across different medical conditions (Kaelbling et al., 1998; Hauskrecht, 2000). Furthermore, real-world constraints such as sensor noise in robotic localization (Thrun et al., 2005), occlusions in autonomous driving (Dosovitskiy et al., 2017), or latent regimes in financial trading (Bacry et al., 2015) mean the agent cannot directly observe the true environmental state. These pervasive sources of ambiguity require agents to maintain beliefs over hidden states and act based on noisy history of observations rather than complete state information.

In this work, we extend the "general agents need world models" theorem from fully observable MDPs to partially observable settings, and then identify a geometrically structured subclass of POMDPs in which the extraction problem admits a sharper formulation. At a high level, we show that goal-optimality in POMDPs forces implicit predictive modeling under episodic resets, that resets are in general necessary for such extraction.

[*]Equal contribution [1]Cambridge Stem Cell Institute, University of Cambridge [2]Hertie Institute for AI in Brain Health, Eberhard-Karls-Universität Tübingen, Germany. Correspondence to: Suraj Yadav <suraj-prakash.yadav@uni-tuebingen.de>, Keshav Goyal <gkeshav91@gmail.com>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

2. Our Contributions

Our contributions are the following.

(i) Extension to POMDPs with approximate resets. We extend the "General Agents Contain World Models" result of Richens et al. (2025) from MDPs to POMDPs. We show that any δ-goal-optimal agent on a sufficiently rich class of finite-horizon goals must implicitly encode the predictive kernel

$$K(h, a) = P(o \mid h, a)$$

of the underlying POMDP (Theorem 4.1). We further prove robustness to approximate resets: if the reset mechanism returns the agent only to an ϵ_h-neighborhood of the target belief state, then the recovered probability obeys

$$|\hat{p} - p| \leq \epsilon_h + O(\delta/\sqrt{n}) + O(1/n).$$

Thus one can trade a controlled linear error in reset quality for a potentially exponential reduction in the number of episodes required (Table 1).

(ii) Necessity of episodic resets. We show that resets are not merely a technical convenience but an information-theoretic prerequisite for extraction in general POMDPs (Theorem 4.2). A simple irreducible three-state counterexample exhibits *coverage collapse*: certain belief states are visited only once and then become unreachable for repeated probing, making the predictive kernel at those states fundamentally unidentifiable from observed behaviour.

(iii) Experimental validation. We validate the theory on synthetic POMDPs using a black-box extraction protocol that observes only the agent's binary goal choices. The experiments confirm the predicted $1/\sqrt{n}$ extraction scaling and the practical tradeoff delivered by approximate resets: exponential savings in episode count at the cost of a linear increase in extraction error.

3. Setup and Definitions

In this section, we formalize the mathematical extension of world-model extraction. We begin by defining the environment; establishing how goals defined over latent states in MDPs are translated into goals over observable histories in POMDPs; the nature of goal-optimal agents.

3.1. Environment: cMDP and POMDP

A **controlled Markov Decision Process** (cMDP) is a tuple $(\mathcal{S}, \mathcal{A}, T)$, where $\mathcal{S} = \{s_1, \ldots, s_n\}$ is a finite state space, $\mathcal{A}$ is a finite action space, and $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the stationary transition kernel with entries $P(s_j \mid s_i, a)$, where $\Delta(\mathcal{S})$ denotes the probability simplex over $\mathcal{S}$.

In a **Partially Observable MDP** (POMDP), the state $s \in \mathcal{S}$ is latent. The agent instead receives observations from a finite set $\mathcal{O}$ via an emission map $E : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{O})$ with entries $P(o \mid s, a)$. The agent acts through the history space $\mathcal{H} = (\mathcal{O} \times \mathcal{A})^*$; we write finite histories as $h \in (\mathcal{O} \times \mathcal{A})^L$ and $h \circ a$ for the concatenation of h with action a. A policy is a map $\pi : \mathcal{H} \to \Delta(\mathcal{A})$, and the central object of interest is the **one-step predictive kernel**

$$K(h, a)(o) \equiv P(o \mid h, a), \quad o \in \mathcal{O}. \tag{1}$$

Belief states. Since s is hidden, the agent maintains a *belief state* $b \in \Delta(\mathcal{S})$, viewed as a row vector in $\mathbb{R}^{1\times|\mathcal{S}|}$, where $b(s) = P(s_t = s \mid h)$. The belief is a sufficient statistic for the history. For fixed $a \in \mathcal{A}$, the transition kernel defines a stochastic matrix $T^a \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ with $T^a_{ij} = P(s_j \mid s_i, a)$, and the emission map defines a diagonal matrix $D^{a,o} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ with $D^{a,o}_{ii} = P(o \mid s_i, a)$. After taking action a and observing o, the belief updates as

$$\tilde{b}' = b\,T^a\,D^{a,o}, \qquad b' = \frac{\tilde{b}'}{\|\tilde{b}'\|_1} \equiv \tau(b, a, o), \tag{2}$$

where $\tau : \Delta(\mathcal{S}) \times \mathcal{A} \times \mathcal{O} \to \Delta(\mathcal{S})$ is the *belief update operator* and $\|\tilde{b}'\|_1 = P(o \mid b, a)$ is the normalising constant. Equivalently, the one-step predictive kernel admits the closed form

$$P(o \mid h, a) = b\,T^a\,E^a\,e_o, \tag{3}$$

where $E^a \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{O}|}$ is the emission matrix with entries $E^a_{ij} = P(o_j \mid s_i, a)$, and $e_o \in \mathbb{R}^{|\mathcal{O}|}$ is the indicator vector for observation o. For a fixed o, the diagonal matrix $D^{a,o} = \mathrm{diag}(E^a\,e_o)$ recovers the per-state emission likelihoods used in the belief update, so that $\tilde{b}' = b\,T^a D^{a,o}$ and $\|\tilde{b}'\|_1 = P(o \mid b, a)$.

3.2. Episodic Resets and History Regeneration

We assume the POMDP is equipped with an **episodic reset** mechanism that returns the environment to the initial belief state b_0 at the start of each trial. The purpose is to reliably revisit the belief state $b(h)$ induced by a target history $h = (o_1, a_1, \ldots, o_L, a_L)$.

We assume the observation sequence $(o_1, \ldots, o_L)$ is reachable with non-zero probability under the corresponding actions. To regenerate $b(h)$, the agent resets to b_0 and executes a fixed **prelude policy** π_{pre} that replays the actions in h, accepting a trial only when the realized observations match $(o_1, \ldots, o_L)$ exactly.

This mechanism is necessary for two reasons. First, since the latent state is unobservable, there is no internal verification that the agent has returned to $b(h)$ without replaying the full history. Second, a general POMDP need not be ergodic:

a single unexpected observation can move the agent to a region of $\Delta(\mathcal{S})$ from which $b(h)$ is unreachable. Episodic resets therefore provide both the guaranteed possibility of return and the verification protocol needed to certify that the target belief state has been reached.

3.3. Goals: From MDP to POMDP

In the fully observable MDP framework, goals are formalized via Linear Temporal Logic (LTL) over the joint state–action space $\mathcal{S} \times \mathcal{A}$ (Pnueli, 1977; Baier and Katoen, 2008). These goals utilize atomic propositions of the form $[(s, a) \in g]$ and temporal operators such as $\bigcirc$ (Next), $\Diamond$ (Eventually), and $\top$ (Now) to express specific task horizons (Pnueli, 1977). Composite goals are formed by the sequencing of these propositions ($\langle \phi_1, \ldots, \phi_L \rangle$) and by disjunctions of such sequences. Within this paradigm, the **depth** of a composite goal is defined as the total number of sequential sub-goals it contains (Richens et al., 2025).

In the Partially Observable (POMDP) setting, however, the latent state $s \in \mathcal{S}$ is not directly accessible to the agent. Consequently, state-based atomic propositions cannot be used for behavioral probing or task specification. To address this, we extend the LTL goal framework by defining atomic propositions over the space of **observable history sequences** $(\mathcal{O} \times \mathcal{A})^*$. In our formulation, a POMDP goal is constructed using observation-action pairs (o, a) (Singh et al., 1994). A goal of depth n in a POMDP requires the agent to achieve n sequential observation-based milestones. This ensures that the agent's behavior is conditioned purely on the information available in its history, allowing for the same rigorous sequencing found in the MDP setting while respecting the constraints of partial observability.

3.4. Goal-Optimality and Bounded Agents

We characterize agents by their ability to satisfy the formal goals defined in the previous section. We define **optimal and bounded goal-conditioned agents** as policies $\pi : \mathcal{H} \to \Delta(\mathcal{A})$ that maximize (or approximately maximize) the probability that a realized trajectory satisfies a given LTL goal ψ.

For a composite goal ψ of depth n, a goal-optimal agent must navigate a sequence of n sub-goals. Formally, an agent is considered **δ-goal optimal** if its success probability for any goal ψ is within a margin δ of the true information-theoretic maximum:

$$P(\pi \models \psi) \geq \max_{\pi'} P(\pi' \models \psi) - \delta. \tag{4}$$

The behavior of such an agent is governed by the relative success probabilities of the available actions. When presented with a choice between two goal branches, ψ_A and ψ_B, a δ-goal optimal agent will consistently select the branch that offers the higher probability of satisfaction, provided the difference in probabilities exceeds the optimality bound δ. This preference-based behavior is the cornerstone of our extraction procedure. By observing which goal an agent pursues when the success probabilities are systematically varied—such as in a binomial switch setup—we can infer the agent's internal estimation of the environment's transition and emission kernels.

For a composite goal of depth n, the agent must not only be optimal for the immediate next observation but must maintain this optimality across all n sequential milestones. This requirement forces the agent to implicitly maintain a sufficient statistic of its history – namely, a world model – to correctly evaluate the success probabilities of its future trajectories.

4. Main Results and Discussion

In this section, we formalize the extraction of the predictive kernel $K(h, a)$ under the realistic constraint that episodic resets may not be perfectly identical. We demonstrate that the error in the recovered probability $\hat{p}$ decomposes linearly into a statistical term derived from the goal-optimality bound δ and a physical term derived from the belief discrepancy ϵ_h. By assuming that the agent is approximately at the same information state whenever a prelude realizes a target history h, we provide a local reset guarantee that avoids the complexities of global error propagation. To ensure the extraction is well-defined, we assume the existence of a per-history neighborhood for resets.

Assumption 3.1 (Per-history local reset). Fix a prelude policy π_{pre} and a realizable history $h \in (\mathcal{O} \times \mathcal{A})^*$. Let $b(h)$ be the nominal belief immediately before applying the probe action after h, assuming the episode began at b_0. We assume there exists $\epsilon_h \geq 0$ such that in every episode that realizes h under π_{pre}, the realized pre-action belief $\hat{b}(h)$ satisfies:

$$TV(\hat{b}(h), b(h)) \leq \epsilon_h. \tag{5}$$

where TV denotes total variation distance. Let $\rho(h) \in (0, 1]$ denote the per-episode probability that π_{pre} realizes h . Let $Y = \mathbb{I}\{o' = o^\star\}$ for a fixed target observation $o^\star \in \mathcal{O}$, and let $Y_1, \ldots, Y_n$ be the indicators from the first n valid trials collected across independent episodes.

Definition 3.2 (Depth-n switch at threshold q). Let $S_n = \sum_{i=1}^{n} Y_i$ and $r = \lfloor nq \rfloor$. We define the paired goals:

$$\Psi_{\leq r} \equiv \{S_n \leq r\} \quad \text{and} \quad \Psi_{>r} \equiv \{S_n > r\}. \tag{6}$$

A δ-goal-optimal agent selects the branch with the larger success probability.

Definition 3.3 (Switch Estimator). For a fixed history h, action a, and outcome $o^\star$, let $p := K(h, a)(o^\star)$ be the true

predictive probability. By sweeping the threshold r and observing the agent's selections, we define the switch index $k^\star := \min\{r : \text{the agent selects } \Psi_{\leq r}\}$, yielding the estimator $\hat{p} := k^\star/n$.

In a POMDP, episodes must first realize the specific sequence h from the initial belief b_0 (Guo et al., 2016). Let $\rho(h) = P(\text{episode realizes } h)$ be the coverage probability. To gather the n valid trials required for extraction, the expected number of attempted episodes is:

$$E[N] = \Theta\left(\frac{n}{\rho(h)}\right). \tag{7}$$

In arbitrary POMDPs, $\rho(h)$ can decay exponentially with the history length $|h|$, leading to an exponential attempt complexity in the worst case (Even-Dar et al., 2005). To mitigate this, we develop an **approximate reset variant** that allows the agent to generate and accept histories h' whose underlying belief states are sufficiently similar to the target belief state $b(h)$ (Kaelbling et al., 1998). By accepting a neighborhood of histories rather than a single exact sequence, we significantly increase the effective coverage probability $\rho(h)$, ensuring that the world-model extraction remains feasible even for longer horizons. This leads to a clean decomposition of the total error into a statistical term $O(\delta/\sqrt{n}) + O(1/n)$ and a reset error term ϵ_h proportional to the belief discrepancy.

Theorem 4.1 (Frequency estimation of $K(h, a)$ under approximate resets)**.** *Let π be a δ-goal optimal agent. Suppose the agent gathers n valid trials using an ϵ_h-approximate reset for history h. Then the indicators Y_i are independent Bernoulli variables with success probabilities p_i satisfying $|p_i - p| \leq \epsilon_h$, and the recovered predictive probability $\hat{p}$ satisfies:*

$$|\hat{p} - p| \leq \underbrace{\epsilon_h}_{\textit{Reset Error}} + \underbrace{O\left(\frac{\delta}{\sqrt{n}}\right) + O\left(\frac{1}{n}\right)}_{\textit{Statistical Error}}. \tag{8}$$

Proof. The complete proof along with the psuedocode of the extraction algorithm is in the appendix A.3. However we explain the strategy of the proof below.

□

4.1. Strategy of the proof

The core theoretical mechanism for world-model extraction relies on the construction of a specific set of composite goals, known as "switch" goals.

We define a **valid** (h, a)**-trial** as a two-stage event within an episode (see Figure 1):

1. The agent successfully realizes the target observable history h from the initial belief b_0.

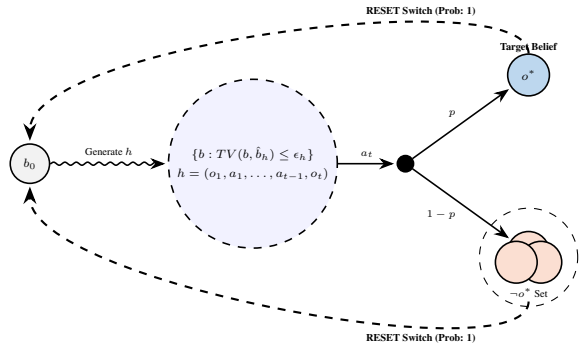

Figure 1. The Switch Mechanism Principle in POMDPs. Upon taking action a_t, the environment branches into the target belief o^* (blue) with probability p or the complement set $\neg o^*$ (orange) with probability $1-p$. The RESET Switch (dashed lines) represents the core assumption of the framework: a guaranteed, universal transition from any outcome back to b_0 with probability 1, bypassing environmental sinks and ensuring the history h can be re-sampled indefinitely for unbiased estimation.

2. The agent executes action a and records the subsequent observation o'.

Each of our switch goals is constructed as a disjunction over sequential sub-goals that (i) recognize history h, (ii) apply action a, and (iii) classify the next observation into success (e.g., observing a target observation o^*) or failure. This process is repeated n times and evaluated against a threshold r, resulting in binary switch goals that compare events of the form $\{S_n \leq r\}$ versus $\{S_n > r\}$, where S_n counts successes (occurrences of o^*) across n valid trials.

Because each switch goal is a disjunction of these sequential sub-goals, the agent's δ-optimal branch choice implements a noisy comparison of binomial tails. Consequently, the agent's choice identifies which side of the threshold r/n the true predictive probability $P(o^*|h, a)$ lies on. This mechanism ensures that the agent's behavior successfully recovers the kernel with an error bound of $O(\delta/\sqrt{n}) + O(1/n)$, effectively extending the MDP readout result ((Richens et al., 2025)) to the POMDP setting .

4.2. Distinction from MDP

In this section we highlight some of the differences with the MDP case.

4.2.1. Object of Recovery

The primary distinction lies in the target of the extraction process. (Richens et al., 2025) recover the state transition kernel $P(s'|s, a)$ of a fully observable MDP. In our setting, we recover the **predictive kernel** $K(h, a) = P(o|h, a)$ over observable histories.

4.2.2. The Failure of Naive Reduction

A common intuition is to attempt a reduction of the POMDP to a fully observable MDP by treating each history as a "state." However, this naive extension fails due to several fundamental obstacles:

- **Non-Returnability of Histories:** In a standard irreducible MDP, an agent can traverse from state s to s' and eventually return to s to collect n independent samples. In a POMDP, any transition from history h to a new observation necessarily increases the history length. Consequently, one cannot simply "return" to history h to sample it again without a full episodic reset to the initial belief b_0 .
- **Belief Space Irreducibility:** Even if we operate in the belief space rather than the history space, we cannot ensure the agent will eventually return to a specific belief state b. Even if the underlying physical environment is irreducible, the belief over histories may not be.
- **The Verification Obstacle:** In a fully observable MDP, the agent can directly observe that it has returned to state s. In a POMDP, implementing the extraction switch requires the agent to verify that the current latent belief is sufficiently close to $b(h)$. However, $b(h)$ is latent; to compute or verify the current belief from observations, one effectively needs to run a belief update filter, which requires knowledge of the transition and observation kernels. Since these kernels are exactly what we are trying to extract, we cannot assume the agent knows when it has "returned" to a specific belief state without already assuming it possesses the world model.

4.3. Why Episodic Resets are Necessary

The preceding extraction results rely fundamentally on episodic resets to the initial belief b_0. A natural question is whether resets are merely convenient or information-theoretically necessary. We show that without resets, even simple, irreducible POMDPs exhibit **coverage collapse** – a phenomenon where certain belief states are visited exactly once and then lost forever, preventing the learning of specific predictive kernels.

Theorem 4.2 (Coverage collapse without resets)**.** *There exists a finite, irreducible POMDP with the following properties:*

1. *The underlying MDP is irreducible and aperiodic.*
2. *For the initial history $h_0 = O_1$ with belief $b(h_0) = (1/2, 1/2, 0)$, there exists $\epsilon_0 > 0$ such that the ball $B_{\epsilon_0}(b(h_0))$ is visited exactly once (at $t = 1$) and never again with probability* 1.
3. *For any observable history h of length L ending in O_1, the stationary frequency satisfies $\rho(h) \leq C \exp(-cL)$ for constants $C, c > 0$.*

Consequently, without resets: (a) learning $p(o|h_0, a)$ is impossible even via interpolation from later observations, and (b) learning $p(o|h, a)$ for longer histories requires $\Omega(\exp(L))$ time steps.

Proof. Proof in the appendix A.3 □

4.3.1. Resets in Theory

The validity of extracting a world model from a goal-optimal agent relies fundamentally on the ability to repeatedly probe the agent's behavior under consistent environmental conditions. In this section we discuss resets and approximate resets along with their practical and theoretical implications. See the appendix A.1 for a detailed discussion.

5. Related Work

The question of whether an agent's behavior implies an internal model touches upon classical identifiability and representation learning (Allman et al., 2009; Hsu et al., 2009). Our work situates itself at the intersection of latent-variable statistics and the theory of optimal control, providing a bridge between predictive modeling and goal-directed agency (Tishby and Polani, 2011). See the Appendix A.2.

6. Experimental Design and Validation

We constructed two test environments with verified properties to evaluate our theoretical claims: **POMDP-10**, consisting of 10 states, 6 observations, and 3 actions, and **POMDP-20**, which contains 20 states, 13 observations, and 3 actions. These environments were engineered to ensure *irreducibility*, where all states are reachable from all other states under some finite action sequence. Furthermore, we ensured *genuine partial observability* by creating multiple confusable state pairs, defined as pairs with a Total Variation (TV) distance < 0.4; specifically, POMDP-10 contains 128 confusable pairs, while POMDP-20 features 546 pairs. The environments also feature a moderate emission entropy, with a normalized entropy within $[0.9, 0.95]$, ensuring that emissions are neither deterministic nor uniform to create realistic partial observability. Transitions follow a structured pattern where action 0 induces forward cycling with a probability of 0.7, action 1 induces backward cycling with a probability of 0.7, and action 2 initiates random jumps with self-loops. To achieve a moderate spread, emissions are drawn from Dirichlet distributions with parameters $\alpha \in [1.5, 3.5]$.

Our experimental protocol decouples the learning and extraction phases to isolate the distinct contributions of each

Table 1. Computational advantage and error scaling for approximate resets across varying ϵ.

ϵ	POMDP-10 ($\lvert h\rvert = 3$, n=200) $\rho(h)$	Att.	Err.	POMDP-20 ($\lvert h\rvert = 3$, n=200) $\rho(h)$	Att.	Err.
0.00	0.004524	25,850	0.0267	0.000534	194,418	0.0165
0.15	0.022283	8,452	0.0259	0.005678	94,754	0.0179
0.20	0.095148	1,565	0.0312	0.016727	12,122	0.0225
0.25	0.236277	757	0.0430	0.068462	2,958	0.0266

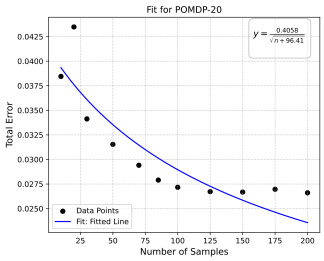

Figure 2. Theoretical fits for POMDP-20.

error source. In the first phase, the agent constructs an internal world model, $\hat{K}$, by collecting $n_{valid} = 100$ valid trials for each history h and action a through episodic resets. This process introduces a fixed learning error floor, which represents the agent's specific competency or inherent inability to fully know the environment; in this state, the agent acts optimally with respect to its own learned beliefs rather than the true environmental transition probabilities. Here, we also compute the realization rate $\rho(h)$ and total attempts required corresponding to each history h. For each POMDP, we generate 61 test histories: one empty history and 15 randomly generated histories at each length $|h| \in \{1, 2, 3, 4\}$, generated by running episodes under random policies.

In the second phase, we treat the learned agent as a black box and apply our switch estimator at varying goal depths $n \in \{10, 20, 30, 50, 70, 85, 100, 125, 150, 175, 200\}$ to extract its internal beliefs. For each n, we compute the switch index $k^* = \min\{r : \Pr(S_n \leq r) \geq \Pr(S_n > r)\}$, where $S_n \sim \text{Bin}(n, \hat{p}_{learned})$, yielding the extracted estimate $\hat{p}_{extracted} = k^*/n$. The extraction algorithm operates without any direct access to the internal kernel $\hat{K}$, relying solely on the observation of the agent's binary choices in response to switch goals, which introduces an extraction error predicted to scale as $O(1/\sqrt{n})$.

6.1. Results

Extraction Error Scaling The extraction error observed in Phase 2 closely follows the theoretical $O(1/\sqrt{n})$ bound established in our analysis (see Figure 2). As shown in Table 2, the extraction error dominates at small values of n but becomes negligible relative to the learning error floor as n increases. This convergence confirms the additive error decomposition predicted by the paper's theoretical framework, demonstrating that our query mechanism effectively recovers the agent's internal state of knowledge as the goal depth increases.

Learning Error (Phase 1) The learning error established in Phase 1, derived from empirical kernel estimation using 100 valid trials, remains constant regardless of the goal depth n utilized in Phase 2. For the **POMDP-10** environment, the mean learning error is 0.032 ± 0.022, while for **POMDP-20**, it is 0.029 ± 0.019. This persistent error floor is a direct consequence of using a finite number of valid trials, making it fundamentally independent of the extraction depth n.

Total Error Decomposition The total error, defined as $|\hat{p}_{extracted} - p_{true}|$, cleanly decomposes into the sum of learning error and extraction error. At lower values of n, the extraction error is the dominant component of the total error; however, at higher values of n, the learning error becomes the primary bottleneck. For instance, in **POMDP-10** at $n = 10$, the total error is 0.0735, but at $n = 200$, it stabilizes at 0.0322 as it asymptotically approaches the learning error floor. These results validate the additive error structure predicted by the paper's theoretical framework, confirming that while extraction error can be minimized through deeper queries, the total accuracy is ultimately limited by the quality of the agent's internal model.

Coverage Collapse The realization rate $\rho(h)$ exhibits exponential decay with history length, confirming the theoretical worst-case analysis. For **POMDP-10**, $\rho(h)$ decreases from 1.0 for an empty history to approximately 0.001 at $|h| = 4$. **POMDP-20** shows even more severe collapse, with $\rho(h) \approx 0.00004$ at $|h| = 4$. This empirical evidence validates the necessity of episodic resets: without such a mechanism, gathering a sufficient number of valid trials for long histories becomes computationally intractable due to the vanishingly small probability of naturally realizing those specific observation sequences.

Approximate Resets and Verifier Advantage To address the coverage collapse, we conducted further experiments using a verifier mechanism to sample approximate histories. In this setting, a history h' is considered valid for a target history h if their resulting belief states are within an ϵ-neighborhood in terms of Total Variation distance. As shown in Table 1, this approach yields a significant compu-

tational advantage by increasing the realization rate, making the extraction from longer histories feasible. However, this gain in efficiency introduces an additional approximation bias; our results demonstrate that the extraction error increases linearly with the tolerance ϵ. This confirms that while approximate resets mitigate the intractability of the coverage collapse, they necessitate a controlled trade-off between computational budget and the precision of the extracted world model.

7. Conclusion

In this paper we demonstrated that a world model—the transition kernel of the underlying environment—can be effectively recovered by observing an agent's binary choices across varying goal depths. Theoretically, we showed that the extraction error scales as $O(1/\sqrt{n})$ and is additive to the agent's own learning error, a decomposition we verified through extensive experiments on synthetic environments. Practically, our introduction of a approximate reset mechanism offers a viable path to overcoming the coverage collapse that typically renders history-based extraction computationally intractable.

A natural next step is to exploit additional structure beyond arbitrary finite POMDPs. In particular, when latent states form a finite group or compact Lie group and observations arise from quotient aliasing $G \to G/H$, the posterior is confined to a single coset and the hidden uncertainty evolves on the subgroup fiber. This suggests sharper extraction procedures that use group structure, better reset mechanisms, and spectral tools adapted to the hidden kernel. We also expect adaptive selection of high-$\rho(h)$ histories and approximate extraction schemes to further reduce sample complexity in large partially observable environments.

References

Allman, E. S., Matias, C., and Rhodes, J. A. (2009). Identifiability of parameters in latent structure models with many observed variables. *Annals of Statistics*, 37(6A):3099–3132.

Azizzadenesheli, K., Lazaric, A., and Anandkumar, A. (2016). Reinforcement learning of POMDPs using spectral methods. *arXiv preprint arXiv:1602.07764*.

Bacry, E., Mastromatteo, I., and Muzy, J.-F. (2015). Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(1):1550005.

Baier, C. and Katoen, J. P. (2008). *Principles of model checking*. MIT Press.

Bellemare, M. G., Naddaf, Y., Veness, J., and Bowling, M. (2013). The arcade learning environment: An evaluation platform for general agents. *Journal of Artificial Intelligence Research*, 47:253–279.

Conant, R. C. and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, 1(2):89–97.

Coumans, E. and Bai, Y. (2016–2021). PyBullet, a Python module for physics simulation for games, robotics and machine learning. Available at: http://pybullet.org.

Dosovitskiy, A., Ros, G., Codevilla, A., Lopez, A., and Koltun, V. (2017). CARLA: An open urban driving simulator. In *Conference on Robot Learning (CoRL)*.

Even-Dar, E., Kakade, S. M., & Mansour, Y. (2005). Reinforcement learning in POMDPs without resets. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 660–665.

Eysenbach, B., Gupta, A., Ibarz, J., and Levine, S. (2019). Diversity is all you need: Learning skills without a reward function. In *International Conference on Learning Representations (ICLR)*.

Florensa, C., Held, D., Geng, X., and Abbeel, P. (2018). Automatic goal generation for reinforcement learning agents. In *International Conference on Machine Learning (ICML)*.

Guez, A., Mirza, M., Gregor, K., Kabra, R., Racanière, S., Weber, T., Raposo, D., Santoro, A., Orseau, L., Eccles, T., Wayne, G., Silver, D., & Lillicrap, T. (2019). An investigation of model-free planning. In *International Conference on Machine Learning (ICML)*, pp. 2464–2473.

Guo, Z. D., Doroudi, S., and Brunskill, E. (2016). A PAC RL algorithm for episodic POMDPs. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*.

Hall, B. C. (2015). *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, 2nd ed. Graduate Texts in Mathematics, Vol. 222. Springer, Cham.

Hauskrecht, M. (2000). Value-function approximations for partially observable Markov decision processes. *Journal of Artificial Intelligence Research*, 13:33–94.

Hsu, D., Kakade, S., and Zhang, T. (2009). A spectral algorithm for learning hidden Markov models. in *Journal of Computer and System Sciences*, 78(5):1460-1480, 2012

Jaques, N., Shen, J., Ghandeharioun, A., Ferguson, C., Lapedriza, A., Jones, N., Gu, S., and Picard, R. (2019). Way off-policy batch deep reinforcement learning of implicit human preferences in dialog. *arXiv preprint arXiv:1907.00456*.

Jin, C., Kakade, S. M., Krishnamurthy, A., & Liu, Q. (2020). Sample-efficient reinforcement learning of undercomplete POMDPs. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 18530–18539).

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1-2):99–134.

Khemakhem, I., Kingma, D., Monti, R., and Hyvarinen, A. (2020). Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 2207–2217.

Lee, J. M. (2013). *Introduction to Smooth Manifolds*, 2nd ed. Graduate Texts in Mathematics, Vol. 218. Springer, New York.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*. American Mathematical Society.

Littman, M. L., Sutton, R. S., and Singh, S. (2001). Predictive representations of state. In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1555–1561.

Liu, M., Zhu, M., and Zhang, W. (2022). Goal-conditioned reinforcement learning: Problems and solutions. *arXiv preprint arXiv:2201.08299*.

Makoviychuk, V., Wawrzyniak, L., Guo, Y., Lu, M., Storey, K., Macklin, M., Hoeller, D., Rudin, N., Allshire, A., Handa, A., and State, G. (2021). Isaac Gym: High performance GPU-based physics simulation for robot learning. *arXiv preprint arXiv:2108.10470*.

Narvekar, S., Peng, B., Leonetti, M., Sinapov, J., Taylor, M. E., and Stone, P. (2020). Curriculum learning for reinforcement learning domains: A framework and survey. *Journal of Machine Learning Research*, 21(181):1–50.

Plappert, M., Andrychowicz, M., Ray, A., McGrew, B., Baker, B., Powell, G., Schneider, J., Tobin, J., Chociej, M., Welinder, P., Kumar, V., and Zaremba, W. (2018). Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*.

Pnueli, A. (1977). The temporal logic of programs. In *18th Annual Symposium on Foundations of Computer Science (SFCS 1977)*, pp. 46–57. IEEE.

Portelas, R., Colas, C., Weng, L., Hofmann, K., and Oudeyer, P.-Y. (2020). Teacher algorithms for curriculum learning of deep RL in continuously parameterized environments. In *Conference on Robot Learning (CoRL)*.

Richens, J., Abel, D., Bellot, A., and Everitt, T. (2025). General agents contain world models. *arXiv:2506.01622*, to appear in *International Conference on Machine Learning (ICML) 2025*.

Singh, S. P., Jaakkola, T., and Jordan, M. I. (1994). Learning without state-estimation in partially observable Markovian decision processes. In *International Conference on Machine Learning (ICML)*, pp. 284–292.

Thrun, S., Burgard, W., and Fox, D. (2005). *Probabilistic Robotics*. MIT Press.

Tishby, N. and Polani, D. (2011). Information theory of decisions and actions. In *Perception-Action Cycle: Models, Architectures, and Hardware*, pp. 601–636. Springer.

Todorov, E., Erez, T., and Tassa, Y. (2012). MuJoCo: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

Young, S., Gasic, M., Thomson, B., and Williams, J. D. (2013). POMDP-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179.

A. Appendix

A.1. Resets in Theory and Practice

A.1.1. Definition of Exact and Approximate Resets

At its most basic level, an episodic reset is a mechanism that returns the environment to a fixed initial configuration or a known initial belief distribution b_0. While exact resets have long been the standard in theoretical frameworks, our work utilizes a more flexible approximate reset variant. This relaxation assumes that when a target history h is realized, the agent's realized belief $\hat{b}(h)$ lies within a Total Variation (TV) neighborhood of the nominal belief $b(h)$, such that $TV(\hat{b}(h), b(h)) \leq \epsilon_h$. This mathematical bridge allows us to maintain theoretical rigor while acknowledging the immense computational difficulty of exact history regeneration in complex POMDPs. By allowing for a small error ϵ_h, we circumvent the need for an exact match of the history string, which would otherwise require a number of samples that scales exponentially with the length of the history.

A.1.2. Alignment with Theoretical Literature

The assumption of episodic resets is well-established in the Reinforcement Learning literature and serves as a primary tool for establishing sample complexity and regret bounds. Theoretical work by Guo, Doroudi, and Brunskill (2016) (Guo et al., 2016) provided a foundational PAC RL algorithm for episodic POMDPs by explicitly assuming that each episode begins from the same initial distribution. Similarly, research by Azizzadenesheli (2016) (Azizzadenesheli et al., 2016) and Jin (2020) (Jin et al., 2020) has utilized episodic structures to derive efficient algorithms for finite-horizon POMDPs. By contrast, studies by Even-Dar (2005) (Even-Dar et al., 2005) on POMDPs without resets have shown that learning in such settings can require sample complexity that is exponential in the horizon. Consequently, our work does not introduce a qualitatively new assumption but rather formalizes the per-history reset structure that underlies the most sample-efficient guarantees in the field.

A.1.3. Practical Applications in Simulation and Robotics

In practice, episodic resets are naturally occurring or explicitly engineered across a variety of domains. Modern robotic simulation platforms, such as MuJoCo (Todorov et al., 2012), Isaac Gym (Makoviychuk et al., 2021), and PyBullet (Coumans and Bai, 2016–2021), as well as classic video game benchmarks like Atari (Bellemare et al., 2013), provide exact state resets by design. Beyond simulation, tasks in localization, navigation, and manipulation are typically structured as repeated trials from controlled starting configurations.

Curriculum learning and staged training offer a particularly sophisticated instance of local resets where a teacher agent can initialize episodes at intermediate checkpoints or "starting configurations" that realize specific history prefixes (Narvekar et al., 2020; Portelas et al., 2020). This allows for targeted probing of the predictive kernel $K(h, a)$ at selected histories of interest, rather than requiring broad coverage over all possible trajectories. In this context, the teacher effectively implements approximate resets to neighborhoods around target belief states $b(h)$.

Even in structured domains where exact resets are difficult to achieve, our method remains robust. As demonstrated in our theoretical analysis, approximate resets satisfying a Total Variation distance bound $TV(\hat{b}(h), b(h)) \leq \epsilon_h$ enable a successful read-out with a bounded additive error. This suggests that practical mechanisms—such as recognizer-based resets that accept episodes based on discriminator confidence, or model-assisted initialization that drives beliefs toward a target—can be utilized to achieve small ϵ_h. Consequently, whether through natural episode boundaries in spoken dialog systems (Young et al., 2013) or generated scenes in autonomous driving (Dosovitskiy et al., 2017), our framework provides a principled way to audit the internal world models of goal-conditioned agents across diverse, real-world applications.

A.1.4. Theoretical Necessity and Coverage Collapse

It is important to emphasize that we do not assume resets merely for analytical convenience; they are an information-theoretic requirement for the extraction of a world model. In this paper, we establish that without a reset-like mechanism, environments are prone to coverage collapse. In these scenarios, the probability of reaching specific history-dependent information states decays exponentially over time, making it impossible to collect the repeated observations necessary to identify the predictive kernel. Resets act as the necessary bridge to circumvent this decay, enabling the repeated sampling required to read out the agent's internal world model.

A.1.5. Limitations and the Boundary of Irreversibility

Despite its utility, the reset assumption defines a clear boundary for the solvability of the world-model extraction problem and is not applicable in strictly irreversible or lifelong learning scenarios. A primary example is medical treatment, where administering a specific drug may permanently alter a patient's physiology; in such cases, one cannot reset the subject to a pre-treatment state to test an alternative action. Similarly, in physical systems where an object is broken or a resource is consumed, the state transition is irreversible. In these single-stream environments, where the agent faces an unbroken sequence of experience without boundaries, coverage collapse ensures that extracting the world model is information-theoretically impossible. Our results thus characterize the specific class of environments where an agent's internal representations can be rigorously extracted.

A.2. Related Work

A.2.1. Identifiability in Latent Variable Models

The identifiability of latent-variable models, such as Hidden Markov Models (HMMs) and POMDPs, has been studied extensively (Allman et al., 2009). Classical approaches, such as spectral methods, provide polynomial-time algorithms for learning HMM parameters under specific rank and separation assumptions (Hsu et al., 2009; Azizzadenesheli et al., 2016). Similarly, researchers have established general identifiability results for finite latent-variable models, demonstrating that distributions over multiple observation sequences can uniquely determine hidden dynamics (Allman et al., 2009). However, these works typically focus on *constructive algorithms*—designing procedures to estimate parameters from data (Hsu et al., 2009). In contrast, our work analyzes the *objective function* itself. We prove that the condition of being goal-optimal fundamentally forces the agent to encode the predictive kernel. This effectively frames identifiability not as a problem to be solved by a specific algorithm, but as an inescapable byproduct of optimal behavior (Richens et al., 2025).

A.2.2. Predictive State Representations (PSRs)

Our choice of the predictive kernel $K(h, a) = P(o|h, a)$ as the object of recovery is deeply rooted in the literature on Predictive State Representations (Littman et al., 2001). PSRs provide a way to model POMDPs without referring to hidden states, instead representing the state as a vector of predictions about the outcomes of future "tests" or sequences of actions and observations (Littman et al., 2001). While the PSR literature focuses on how an agent can *learn* these representations to improve planning, we show that any agent capable of achieving a sufficiently diverse set of goals must *already* contain such a representation (Tishby and Polani, 2011). This connects the operational success of goal-conditioned policies to the underlying predictive dynamics of the environment.

A.2.3. The Good Regulator and Internal Models

The idea that a controller must reflect the system it controls is a cornerstone of cybernetics, notably formalized by the Good Regulator Theorem (Conant and Ashby, 1970). This theorem posits that any system that is effective at regulation must be a model of that system (Conant and Ashby, 1970). While conceptually profound, the original formulation has been criticized for being too broad. Recent work in the fully observable MDP setting provided a modern, rigorous version of this idea by showing that goal-optimality forces model learning (Richens et al., 2025). Our work extends this philosophical lineage to the POMDP space, providing the necessary mathematical framework to handle the ambiguity inherent in partial observations.

A.2.4. Distinction: Objective Analysis vs. Algorithm Design

A key distinction of our contribution is its focus on objective-based necessity rather than algorithmic design (Allman et al., 2009). In standard POMDP research, the goal is often to design a training algorithm (such as an RL agent or a spectral learner) that converges to a world model (Azizzadenesheli et al., 2016; Jin et al., 2020). Our work takes the agent's optimality as a given and investigates what information is information-theoretically required to maintain that optimality. We do not propose a new training regime; instead, we prove that any agent that succeeds at achieving a rich set of history-based goals has already solved the identifiability problem internally. This perspective complements Inverse Reinforcement Learning (IRL); while IRL uses a policy and a world model to identify a goal, we use a policy and a goal to identify the world model (Jaques et al., 2019).

A.2.5. Representation Learning and Goal-Conditioned Policies

Finally, our work relates to the field of representation learning, which investigates how agents compress high-dimensional histories into useful latent features (Khemakhem et al., 2020). Goal-conditioned policies are often touted for their ability to learn these features implicitly without the need for an explicit transition model (Liu et al., 2022; Plappert et al., 2018). By proving that a predictive kernel is necessary for optimality, we provide a theoretical justification for why these "model-free" agents often exhibit emergent planning-like behaviors (Guez et al., 2019). We show that the "features" learned by an optimal agent must, at a minimum, be sufficient to reconstruct the predictive dynamics of the environment.

A.3. Proof of the Theorems

We now give a proof of the Theorem 4.1

Proof. We break the proof in a few steps.

1. Bounding the Reset Error: In each trial i, the environment resets to a belief state $\hat{b}_i(h)$ such that $TV(\hat{b}_i(h), b(h)) \leq \epsilon_h$. Let $p_i := P(o^*|\hat{b}_i(h), a)$. Since the predictive map $b \mapsto P(o^*|b, a)$ is 1-Lipschitz with respect to Total Variation, we have $|p_i - p| \leq \epsilon_h$. Averaging over n trials, the mean success probability $\bar{p} = \frac{1}{n}\sum p_i$ satisfies:

$$|\bar{p} - p| \leq \epsilon_h. \tag{9}$$

2. The Difference Function and Local Limit Theorem: We recover $\hat{p}$ via the preference discriminator function $D(r)$, defined as the difference in success probabilities between the branch $\Psi_{>r}$ and $\Psi_{\leq r}$:

$$D(r) := \Pr(S_n > r) - \Pr(S_n \leq r) = 1 - 2\Pr(S_n \leq r).$$

Since the CDF is non-decreasing, $D(r)$ is non-increasing. Crucially, the step size of $D(r)$ is given by:

$$\begin{aligned} D(r-1) - D(r) &= (1 - 2\Pr(S_n \leq r-1)) \\ &\quad - (1 - 2\Pr(S_n \leq r)) \\ &= 2\Pr(S_n = r). \end{aligned} \tag{10}$$

Assuming the variance is non-degenerate (i.e., $\sum p_i(1-p_i) = \Theta(n)$), the Local Limit Theorem for Poisson-Binomial sums guarantees that for t in a $\sqrt{n}$-neighborhood of the mean μ:

$$\Pr(S_n = t) \geq \frac{c_0}{\sqrt{n}},$$

for some constant $c_0 > 0$. Substituting this into Eq. 10 implies that $D(r)$ decreases by at least $2c_0/\sqrt{n}$ per step in the central region.

3. The Optimality Switch Band: The switch index is defined as $k^* := \min\{r : \text{Agent selects } \Psi_{\leq r}\}$.

- At $r = k^*$, the agent selects $\Psi_{\leq k^*}$. By δ-optimality, $\Pr(S_n \leq k^*) \geq \Pr(S_n > k^*) - \delta$, which implies $D(k^*) \leq \delta$.
- At $r = k^* - 1$, the agent selects $\Psi_{>k^*-1}$ (by minimality of k^*). By δ-optimality, $\Pr(S_n > k^* - 1) \geq \Pr(S_n \leq k^* - 1) - \delta$, which implies $D(k^* - 1) \geq -\delta$.

Thus, the switch occurs exactly where $D(r)$ crosses the $[-\delta, \delta]$ band:

$$-\delta \leq D(k^* - 1) \quad \text{and} \quad D(k^*) \leq \delta. \tag{11}$$

4. Convergence: Let m be the median of S_n (where $D(m) \approx 0$). Since $D(r)$ decreases linearly at rate $\Theta(1/\sqrt{n})$ and the switch occurs within a window of height 2δ, the horizontal distance from the median is bounded by:

$$|k^* - m| \leq \frac{2\delta}{2c_0/\sqrt{n}} + O(1) = O(\delta\sqrt{n}).$$

Using the property that $|\mu - m| < 1$ for Poisson-Binomial distributions, we have $|k^* - \mu| \leq O(\delta\sqrt{n})$. Dividing by n yields the statistical error $|\hat{p} - \bar{p}| \leq O(\delta/\sqrt{n}) + O(1/n)$. Combining this with Step 1 completes the proof. □

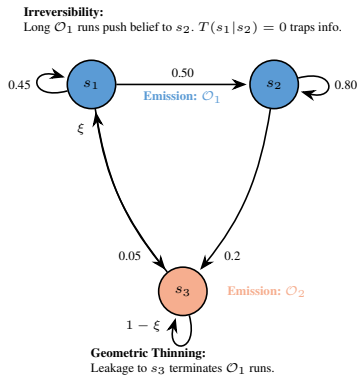

Figure 3. Structural counterexample illustrating the fundamental constraints on world model extraction. The blue states (s_1, s_2) emit identical observations $\mathcal{O}_1$, while the orange state (s_3) emits $\mathcal{O}_2$. The transition dynamics demonstrate Part (ii): Irreversibility, where long sequences of $\mathcal{O}_1$ shift the agent's belief into the s_2 trap from which s_1 is unreachable, and Part (iii): Geometric Thinning, where leakage to s_3 at rates ϵ and 0.05 ensures that the probability of realizing specific history sequences $\rho(h)$ decays exponentially.

Proof. of Theorem 4.2. We construct a 3-state POMDP with states $\{s_1, s_2, s_3\}$ and observations $\{O_1, O_2\}$. Emissions are deterministic:

$$E(O_1 \mid s_1) = E(O_1 \mid s_2) = 1, \qquad E(O_2 \mid s_3) = 1.$$

The transition kernel (identical for both actions) is

$$\begin{aligned} T(\cdot \mid s_1) &= (0.45,\ 0.50,\ 0.05), \\ T(\cdot \mid s_2) &= (0.00,\ 0.80,\ 0.20), \\ T(\cdot \mid s_3) &= (\xi,\ 0,\ 1-\xi), \qquad \xi \in (0,1). \end{aligned}$$

The induced Markov chain on states is irreducible: from s_1 one reaches s_2 and s_3 in one step; from s_2 one reaches s_3 in one step; from s_3 one reaches s_1 in one step with probability ξ, and from s_1 one reaches s_2 in one step. Hence every state communicates with every other.

Belief dynamics on the O_1-face. Whenever the observation is O_1, the posterior assigns zero mass to s_3. Thus beliefs on the O_1-face have the form

$$b = (1-x,\ x,\ 0), \qquad x := \Pr(s_2 \mid h) \in [0,1].$$

Let F denote the Bayes update map on this face conditional on observing O_1 one more time. Starting from $b = (1-x, x, 0)$, the one-step predictive distribution over next state is

$$bT = \big(0.45(1-x),\ 0.50(1-x) + 0.80x,\ 0.05(1-x) + 0.20x\big).$$

Conditioning on observing O_1 (i.e. next state in $\{s_1, s_2\}$) renormalizes the first two coordinates, so the updated coordinate $x' = \Pr(s_2 \mid hO_1)$ is

$$F(x) = \frac{0.50(1-x) + 0.80x}{0.45(1-x) + 0.50(1-x) + 0.80x} = \frac{0.50 + 0.30x}{0.95 - 0.15x}. \tag{12}$$

A direct calculation gives $F(0) = \frac{10}{19} \approx 0.5263$ and $F(x) > x$ for all $x \in (0,1)$, hence the sequence $x_{t+1} = F(x_t)$ is strictly increasing and converges to 1 whenever it is defined (i.e. along runs of O_1).

Re-entry after observing O_2. Observing O_2 implies the latent state is s_3 with probability 1, so immediately after an O_2 observation the belief is δ_{s_3}. On the next step, the chain transitions from s_3 to s_1 with probability ξ and remains in s_3 with probability $1-\xi$. Consequently, the first time the process returns to observing O_1 after an O_2 run, the latent state must be s_1 (since s_2 is unreachable from s_3 in one step), and therefore the belief upon this return to the O_1-face is exactly

$$b = (1, 0, 0) \quad \text{equivalently } x = 0.$$

Thus every O_2-excursion followed by a return to O_1 resets the face-coordinate to $x = 0$, and the next O_1 update moves it deterministically to $x = F(0) = 10/19$.

Table 2. Extraction error vs. goal depth n.

n	POMDP-10 Total Error	POMDP-10 Extract Error	POMDP-20 Total Error	POMDP-20 Extract Error
10	0.0735 ± 0.0469	0.0641 ± 0.0287	0.0384 ± 0.0046	0.0394 ± 0.0286
20	0.0489 ± 0.0295	0.0342 ± 0.0143	0.0435 ± 0.0223	0.0317 ± 0.0158
30	0.0430 ± 0.0295	0.0220 ± 0.0101	0.0341 ± 0.0229	0.0209 ± 0.0099
50	0.0380 ± 0.0246	0.0142 ± 0.0062	0.0315 ± 0.0199	0.0136 ± 0.0055
70	0.0360 ± 0.0233	0.0107 ± 0.0041	0.0294 ± 0.0180	0.0093 ± 0.0045
85	0.0351 ± 0.0246	0.0070 ± 0.0036	0.0279 ± 0.0186	0.0074 ± 0.0036
100	0.0343 ± 0.0229	0.0051 ± 0.0024	0.0272 ± 0.0201	0.0063 ± 0.0029
125	0.0339 ± 0.0219	0.0054 ± 0.0021	0.0267 ± 0.0194	0.0053 ± 0.0024
150	0.0327 ± 0.0224	0.0037 ± 0.0019	0.0267 ± 0.0187	0.0045 ± 0.0016
175	0.0330 ± 0.0224	0.0034 ± 0.0017	0.0270 ± 0.0191	0.0035 ± 0.0017
200	0.0322 ± 0.0226	0.0030 ± 0.0014	0.0266 ± 0.0194	0.0033 ± 0.0015

Table 3. Coverage collapse: mean realization rate $\rho(h)$ and episode attempts by history length.

$\|h\|$	POMDP-10 $\rho(h)$	POMDP-10 Attempts	POMDP-20 $\rho(h)$	POMDP-20 Attempts
0	1.000	100	1.000	100
1	0.171	611	0.075	1,359
2	0.035	3,338	0.006	17,234
3	0.005	23,384	0.0005	2,02,450
4	0.0008	1,38,632	0.00004	10,00,000.0

Transient neighbourhood of the initial belief. Fix the initial belief on the O_1-face to be $b_0 = (1/2, 1/2, 0)$, i.e. $x_0 = 1/2$. Let

$$\eta := F(0) - x_0 = \frac{10}{19} - \frac{1}{2} = \frac{1}{38}.$$

Then after any return to the O_1-face from an O_2-excursion, the very next O_1 observation forces x to equal $F(0) = x_0 + \eta$. Moreover, along any run of consecutive O_1 observations we have $x_{t+1} = F(x_t) > x_t$, so once $x \geq x_0 + \eta$ it remains $\geq x_0 + \eta$ until the next O_2 occurs. Therefore the open interval $(x_0 - \eta,\, x_0 + \eta)$ is visited at time $t = 0$ but never again: after the first return map to the O_1-face the coordinate jumps to $x_0 + \eta$ and thereafter drifts upward away from x_0. In this precise sense, the neighbourhood $B_\eta(x_0)$ of the information state is transient.

Exponential decay of long histories (geometric thinning). On the O_1-face with coordinate x, the conditional probability of observing O_2 at the next step equals the predictive probability of transitioning to s_3:

$$\Pr(O_2 \mid x) = 0.05(1 - x) + 0.20x = 0.05 + 0.15x \;\geq\; 0.05.$$

Hence, at every step during an O_1-run there is at least a 0.05 chance to terminate the run by seeing O_2. It follows that the length R of any consecutive run of O_1 observations is stochastically dominated by a geometric random variable with success parameter 0.05:

$$\Pr(R \geq L) \leq (0.95)^{L-1}.$$

In particular, for any fixed length-L history h that lies entirely on the O_1-face (i.e. consists of L consecutive O_1 observations with the corresponding action sequence), its realization rate satisfies

$$\rho(h) \;\leq\; \Pr(R \geq L) \;\leq\; (0.95)^{L-1}.$$

This proves an exponential decay upper bound of the claimed form. □

Algorithm 1 Execute Prelude Verification

Require: Initial belief state b_0
Require: Target observable history $h = (o_1, a_1, \ldots, o_L, a_L)$
Require: Prelude policy π_{pre}

```
0: function EXECUTE_PRELUDE(b_0, h, π_pre)
0:   loop
0:     Reset environment to initial belief b_0
0:     observed_sequence ← []
0:     for step = 1 to L do
0:       Take action a_step using π_pre
0:       Record resulting observation o_step
0:       Append o_step to observed_sequence
0:     end for
0:     if observed_sequence == (o_1, ..., o_L) then
0:       return SUCCESS {History h is successfully realized}
0:     else
0:       continue loop {History diverged, discard attempt}
0:     end if
0:   end loop
0: end function=0
```

Algorithm 2 Estimate Predictive Kernel $\hat{K}(h, a)(o^*)$

Require: Goal-conditioned policy $\pi(a_t|h_t; \psi)$
Require: Initial belief state b_0
Require: Target observable history h, target action a, target observation o^*
Require: Precision parameter $n \in \mathbb{N}$, alternative action $b \neq a$
Require: Prelude policy π_{pre}

```
0: function ESTIMATE_PREDICTIVE_KERNEL(π, b_0, h, a, o*, n, b)
0:   k* ← n
0:   for k = 1 to n do
0:     φ_0 ← [A_0 = a]
0:     φ'_0 ← [A_0 = b]
0:     φ_1 ← ◇[EXECUTE_PRELUDE(b_0, h, π_pre) is SUCCESS ∧ A = a]
0:     φ_2 ← ◯[O = o*]
0:     φ'_2 ← ◯[O ≠ o*]
0:     ψ_fail ← ⟨φ_1, φ'_2⟩
0:     ψ_success ← ⟨φ_1, φ_2⟩
0:     ψ_a(k, n) ← ⋁ sequences with r ≤ k successes of ⟨φ_0, (ψ_fail ∨ ψ_success)×n⟩
0:     ψ_b(k, n) ← ⋁ sequences with r > k successes of ⟨φ'_0, (ψ_fail ∨ ψ_success)×n⟩
0:     ψ_ab(k, n) ← ψ_a(k, n) ∨ ψ_b(k, n)
0:     a_0 ← π(a_0|b_0; ψ_ab(k, n))
0:     if a_0 == a then
0:       k* ← k
0:       break
0:     end if
0:   end for
0:   p̂ ← k*/n
0:   return p̂
0: end function=0
```