COUNTERFACTUALS WITHOUT RESETS: IDENTIFYING POMDPS

Anonymous authors
Paper under double-blind review

ABSTRACT

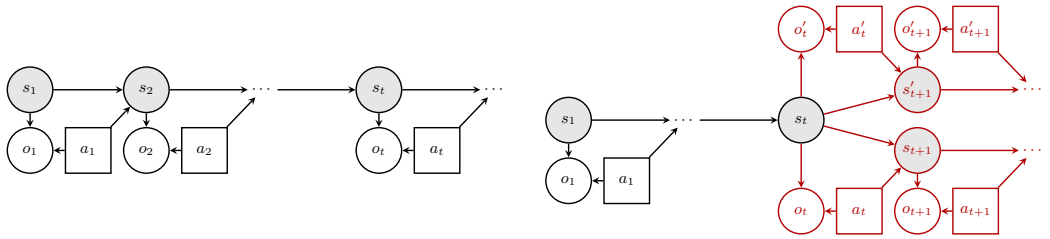
Many problems require estimating a world-dynamics model from interaction data where the observations do not uniquely reveal the underlying state. Our goal is to estimate time-homogenous discrete Partially Observable Markov Decision Processes (POMDPs) from a single trajectory of action and observations without any resets. We find that access to *counterfactual samples* provides sufficient conditions to learn a larger class of POMDPs. Counterfactual samples are samples from the joint distribution of all possible observations sequences obtained from executing multiple action sequences from the same state. In general, when gathering data from a single trajectory, it is not possible to gather counterfactual samples for a partially observable system. We therefore restrict our attention to the class of POMDPs that can be described as deterministic finite-state machines corrupted with bounded transition and observation noise. We show that deterministic machines contain *anchor sequences*, which can fully determine the underlying state when observed. We provide an algorithm, Counterfactual Successive Projection (CfactSP), that uses ideas from tensor decomposition and nonnegative matrix factorization to leverage these anchor sequences to identify the POMDP from a single reset-free data sequence. CfactSP is provably robust to a bounded number of corruptions in transition and observations. Empirical results suggest sample-complexity is substantially lower in practice than the guaranteed bound.

1 INTRODUCTION

Animals and humans learn the world structure from a long, uninterrupted trajectory of actions and observations. We investigate algorithmic approaches to do the same. Our goal is to design an algorithm to estimate a POMDP from a single sequence of actions and observations without reset, without ever observing the latent system state or knowing the number of discrete states. This objective is more stringent than many approaches described to 'learn' partially observable systems, which is to learn some predictive model of the future given a history of observations and actions (Littman & Sutton, 2001; Ha & Schmidhuber, 2018; Allen et al., 2024). In prior work, constructing explicit state spaces allows for swapping a reward function to change planning behavior without relearning the whole model (Shaw et al., 2026). This identification problem is also interesting in its own right, and variants have been studied in automatic control (Anderson, 1999; Vidyasagar, 2011), natural language processing (Balle et al., 2014), and reinforcement learning (Azizzadenesheli et al., 2016; Guo et al., 2016) under many different technical restrictions.

The first step of our study investigates the conditions under which a broader class of POMDPs can be identified from a single trajectory without reset, namely, the set of POMDPs where transitions and observations are time-invariant. Our first contribution is to show that to identify this class of POMDPs, that collecting *counterfactual samples* (Pearl, 2000), that is, samples of multiple observation sequences from multiple action sequences executed from the *same state* (fig. 1b) can help relax identification assumptions from prior work (Azizzadenesheli et al., 2016; Guo et al., 2016). Only after collecting a sufficient number of these trajectory sets can an algorithm recover the number of hidden states for all POMDPs in this class. However, an agent can only collect a counterfactual sample if it can attempt one sequence of actions, obtain the associated observations, *reset to the starting state*, and try another sequence of actions to obtain a new set of observations. Since the state is partially observed, the agent cannot determine whether it has returned to the same starting state to try a different sequence of actions. This intuition is substantiated by prior work that shows that it

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.



(a) POMDP (b) Counterfactual sample

Figure 1: Directed graphical models of a POMDP and a counterfactual sample consisting of two tests from the same state s_t (Pearl, 2000). Latent states are shaded. In the twin graph, the nodes that describe the distribution of two counterfactual samples are shaded in red.

is *impossible* to identify this class of discrete time-homogenous POMDPs from a single trajectory without reset (Shaw et al., 2026).

Instead, we choose to operate in a restricted setting that we believe captures much of the physical world: a POMDP that is deterministic but is subject to low-likelihood corruptions of the transitions and observations. Deterministic POMDPs, under a mild connectivity assumption, have a natural state reset property, which we call *anchor sequences*: action-observation sequences, that, when observed, fully determine the latent state. Whenever the same anchor sequence is observed more than once in a single stream of data, the steps that follow each occurrence begin at the same latent state, and so may be folded into a counterfactual sample. Our second contribution, an algorithm we call counterfactual successive projection (cfactSP), draws on ideas from tensor decomposition (Bhaskara et al., 2014b) and nonnegative matrix factorization (Arora et al., 2013) to exploit these anchor sequences to identify uncorrupted POMDPs without knowing the sequences in advance. We also prove that, with high probability, cfactSP recovers the uncorrupted POMDP under a bounded amount of corruption noise. Our numerical simulations suggest that our algorithm performs better in practice than our theoretical results would suggest.

2 Problem Setting: POMDP Learning Without Reset

We model the environment as a discrete POMDP $(b_0, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{Z})$, where $\mathcal{S} = \{s^1, \ldots, s^r\}$, $\mathcal{A} = \{a^1, \ldots, a^n\}$, and $\mathcal{O} = \{o^1, \ldots, o^m\}$ are finite alphabets of states, actions, and observations, respectively. Reward and discount factors are omitted because the focus of our study is *not* to find an optimal policy. The distribution b_0 is the initial state distribution of the POMDP. The transition dynamics likelihoods $\mathcal{T}$ and observation emission likelihoods $\mathcal{O}$ are expressed in terms of row-stochastic matrices $\mathcal{T} = \{T^a, a \in \mathcal{A}\}$, where $T^a_{ij} = \mathrm{P}(s_{t+1} = s^j \mid s_t = s^i, a_t = a)$, and $\mathcal{Z} = \{O^a : a \in \mathcal{A}\}$, where $O^a_{ij} = \mathrm{P}(o_t = o^j \mid s_t = s^i, a_t = a)$. Lastly, we assume that the POMDP is expressed in a *minimal form*, which states that for each pair of states, $s^i, s^j \in \mathcal{S}$, there exists $a \in \mathcal{A}$ such that either $T^a_{i,:} \neq T^a_{j,:}$ or $O^a_{i,:} \neq O^a_{j,:}$.

In general, upper-indices will refer to elements of sets $\mathcal{A}$, $\mathcal{O}$, and $\mathcal{S}$ (e.g. $\mathcal{S} = \{s^1, \ldots, s^r\}$) and lower indices will refer to time indexing (e.g. $s_1, s_2, \ldots, s_t, s_{t+1}$ for $t > 0$ etc.). The symbol λ will be used for the empty string. We use P for the probability law of the POMDP, whose arguments are events (e.g., $\mathrm{P}(s_{t+1} = s^j \mid s_t = s^i, a_t = a)$), and p for its likelihood (mass function), whose arguments are outcomes (e.g., $p(s_{t+1} \mid s_t, a_t)$, read as a function of the values taken by s_t, a_t, and s_{t+1}). Following Littman & Sutton (2001), we refer to action-observation sequence $(ao)_{1:t}$ that were seen in the past (relative to a time t) a *history*, and an action-observation sequence $(ao)_{t+1:T}$ that the may be seen in the future as a *test*. These strings are analogous to *prefixes* and *suffixes* in natural language processing (Balle et al., 2014).

We consider a POMDP induction problem of the following form. Given: (1) finite action and observation alphabets $\mathcal{A} = \{a^1, \ldots, a^n\}$ and $\mathcal{O} = \{o^1, \ldots, o^m\}$ and (2) a single observation string $(a_1, o_1, \ldots, a_N, o_N)$, drawn by selecting actions a_t uniformly at random from $\mathcal{A}$ and recording the resulting observations o_t according to the ground truth POMDP, we wish to compute a POMDP

model $(\hat{b}_0, \hat{\mathcal{S}}, \mathcal{A}, \mathcal{O}, \hat{\mathcal{T}} = \{\hat{T}^a : a \in \mathcal{A}\}, \hat{\mathcal{Z}} = \{\hat{O}^a : a \in \mathcal{A}\})$ such that the transition dynamics and observation emissions match the true model parameters up to a relabelling of state indices. Formally, there exists a permutation mapping $\sigma : \mathcal{S} \to \hat{\mathcal{S}}$ such that $T^a_{i,j} = \hat{T}^a_{\sigma(i),\sigma(j)}$ and $O^a_{i,j} = \hat{O}^a_{\sigma(i),j}$. Critically, this means that the algorithm must determine the number of states $|\mathcal{S}|$ which we assume is *not* provided as part of the input.

We examine the conditions under which the algorithm can recover the ground truth model. Specifically, we will focus on a special case of this problem where the data are generated by interaction with a *deterministic* POMDP but corrupted by a small amount of noise.

3 Identifying POMDPs from Counterfactuals

This section discusses whether it is possible to identify a time-homogenous POMDP from a single trajectory. To do so, we must first introduce some machinery used to study this problem (Vidyasagar, 2011; Singh et al., 2012; Anandkumar et al., 2012). From a sequence of actions and observations, a place to begin inferring the model parameters is to track co-occurences between past histories and future tests. We organize these co-occurences into (a set of) arrays, one dimension indexed by histories and the other indexed by tests.

Definition 1 (System Dynamics Matrix (Singh et al., 2012))**.** *A* System Dynamics Matrix (SDM) *of lengths K and L for $K, L > 0$, notated as $\mathcal{M}^{K,L}$, is a matrix with $\left(\sum_{i=1}^{K} |\mathcal{A}|^i |\mathcal{O}|^i + 1\right)$ rows and $\left(|\mathcal{A}|^L |\mathcal{O}|^L\right)$ columns, whose rows are indexed by* histories *of lengths 0 (null string) to K and columns are indexed by* tests *of length L. An entry of the SDM indexed by history $(a'o')_{1:t}$ and test $(a'o')_{t+1:t+L}$ is defined as*

$$\mathcal{M}^{K,L}_{(a'o')_{1:t},(a'o')_{t+1:t+L}} = \mathrm{P}(o_{t+1:t+L} = o'_{t+1:t+L} \mid (ao)_{1:t} = (a'o')_{1:t}, a_{t+1:t+L} = a'_{t+1:t+L}). \tag{1}$$

We refer to infinite set of matrices $\mathcal{M}^{K,L}$ for all $K, L > 0$ as $\mathcal{M} = \{\mathcal{M}^{K,L} : K, L > 0\}$. Furthermore, we define the rank *of $\mathcal{M}$ to be* $\mathrm{rank}(\mathcal{M}) = \sup_{K,L>0}\{\mathrm{rank}(\mathcal{M}^{K,L})\}$.

Informally, we often think of the set of matrices $\mathcal{M}$ as a single infinite-by-infinite matrix, from which we take a finite *submatrix* for computation with history and test lengths K and L to be inferred from context. This section focuses on the *identifiability conditions* of POMDPs from action-observation trajectories, so we assume we are operating working with an exact SDM as defined in eq. (1).

Given a history $(ao)_{1:K}$ and a test $(ao)_{K+1:K+L}$, an SDM can be expressed as the multiplication of two nonnegative matrices, $\mathbf{Forw}^K$ and $\mathbf{Back}^L$. Examining a single entry, we observe

$$\begin{aligned}
\mathcal{M}^{K,L}_{(ao)_{1:K},(ao)_{K+1:K+L}} &= p(o_{K+1:K+L} \mid (ao)_{1:K}, a_{K+1:K+L}) \\
&= \sum_{s_{K+1} \in \{s^1,\ldots,s^r\}} p(o_{K+1:K+L}, s_{K+1} \mid (ao)_{1:K}, a_{K+1:K+L}) \\
&= \sum_{s_{K+1} \in \{s^1,\ldots,s^r\}} \underbrace{p(o_{K+1:K+L} \mid s_{K+1}, a_{K+1:K+L})}_{\mathbf{Back}\text{ entry}} \underbrace{p(s_{K+1} \mid (ao)_{1:K})}_{\mathbf{Forw}\text{ entry}}.
\end{aligned} \tag{2}$$

The second line unmarginalizes the state reached after the history, and the third applies the fact that S_{K+1} d-separates $(ao)_{K+1:K+L}$ from $(ao)_{1:K}$ in the graphical model in fig. 1a. We can interpret the third line (eq. (2)) as an inner-product, which allows us to vectorize this calculation. We call $\mathbf{Forw}^K$ a *forward matrix*, whose *row corresponding to history* $(a'o')_{1:t}$ encodes the distribution over possible states having observed $(a'o')_{1:t}$, e.g.,

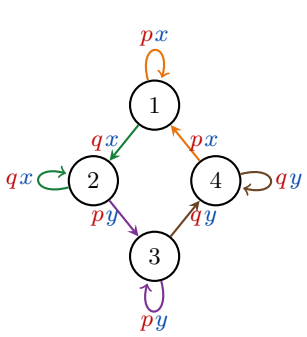
$$\mathbf{Forw}^K_{(a'o')_{1:t},i} = \mathrm{P}(s_{t+1} = s^i \mid (ao)_{1:t} = (a'o')_{1:t}).$$

We call $\mathbf{Back}^L$ a *backward matrix*, whose *column corresponding to a test* $(a'o')_{t+1:t+L}$ encodes the likelihood of that test being observed conditioned on *starting* from a particular state, e.g.,

$$\mathbf{Back}^L_{i,(a'o')_{t+1:t+L}} = \mathrm{P}(o_{t+1:t+L} = o'_{t+1:t+L} \mid s_{t+1} = s^i, a_{t+1:t+L} = a'_{t+1:t+L}). \tag{3}$$

Multiplying these matrices together recovers the $\mathcal{M}^{K,L}$ by evaluating eq. (2) for every combination of history indexing **Forw** and test indexing **Back**. In this paper, we make the assumption that **Forw** is full-rank. Many algorithms, including the one presented in this paper, attempt to learn POMDPs by constructing this decomposition (or a related factorization) first (Wolfe et al., 2005; Boots et al., 2011; Balle et al., 2014; Shaw et al., 2026).

While it is tempting to factor $\mathcal{M}^{K,L}$ first via some rank-decomposition technique (Boots et al., 2011; Balle et al., 2014; Shaw et al., 2026) to identify the number of states, even if **Forw** is full-rank, we cannot. The issue is that $\mathrm{rank}(\mathcal{M})$ is only a *lower bound* of the number of states of the *minimal POMDP*, which can be strict, because **Back** can be rank-deficient (see running example below). Factoring this way can produce predictive models (Singh et al., 2012; Boots et al., 2011), but does not reveal the minimum number of states of the POMDP for identification. Another issue, even with $\mathrm{rank}(\mathcal{M}) = |\mathcal{S}|$, is that there exist families of POMDPs that yield the same SDM, so multiple valid factorings may occur. While predictive models may be good to learn a single reward function for planning, swapping the reward would require re-learning the model (Izadi & Precup, 2008) or the corresponding transition and observation models (Shaw et al., 2026).



(a) State transitions.

Forw^2

h	s^1	s^2	s^3	s^4
λ	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$	$\frac{1}{4}$
px	**1**	0	0	0
qx	0	**1**	0	0
py	0	0	**1**	0
qy	0	0	0	**1**
$pxqx$	0	**1**	0	0
$qypx$	**1**	0	0	0
⋮	⋮	⋮	⋮	⋮

$\times$

Back^1

	px	py	qx	qy
s^1	**1**	0	**1**	0
s^2	0	**1**	**1**	0
s^3	0	**1**	0	**1**
s^4	**1**	0	0	**1**

$=$

$\mathcal{M}^{2,1}$

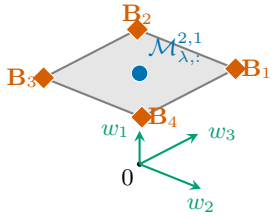
h	px	py	qx	qy
λ	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
px	**1**	0	**1**	0
qx	0	**1**	**1**	0
py	0	**1**	0	**1**
qy	**1**	0	0	**1**
$pxqx$	0	**1**	**1**	0
$qypx$	**1**	0	**1**	0
⋮	⋮	⋮	⋮	⋮

(b) SDM factorization.

$\mathsf{cfactBack}(2)^1 = \left(\mathbf{Back}^1\right)^{\odot 2}$

	$pxqx$	$pypy$	$pxpx$	$pypy$	$qxqx$	$qxqx$	$qypx$	$qypy$	$qxqy$	$qxqy$	$qypy$	$qxqx$	$qypx$	$qypy$	$qxqy$	$qxqy$
s^1	**1**	0	0	0	1	0	0	0	1	0	1	0	0	0	0	0
s^2	0	**1**	0	0	0	0	0	1	0	1	1	0	0	0	0	0
s^3	0	0	**1**	0	0	0	0	1	0	0	0	0	0	1	0	1
s^4	0	0	0	**1**	1	0	0	0	0	0	0	0	1	0	0	1

(c) Counterfactual backward matrix.



(d) Row geometry.

Figure 2: Merry-Go-Round has four states and an SDM of rank three. (a) Transition labels give the action observation. (b) Selected histories and one-step tests in $\mathbf{Forw}^2 \cdot \mathbf{Back}^1 = \mathcal{M}^{2,1}$; λ denotes the empty history. Anchor histories select individual rows of $\mathbf{Back}^1$, copying them into the SDM. (c) The $\mathsf{cfactBack}(2)^1$ matrix, in which each entry multiplies two test probabilities from the *same starting state*; the four shaded columns form I_4. (d) The backward rows $\mathbf{B}_i = \mathbf{Back}^1_{i,:}$ form a square, shown in coordinates of their three-dimensional linear span. SDM rows are convex combinations of these vertices.

A running example. We introduce the deterministic POMDP *Merry-Go-Round* depicted in fig. 2a, consisting of four states, two actions $\{p, q\}$ and two observations $\{x, y\}$. This POMDP was constructed so that the inference over the state the agent came from before observing a particular *test* would be ambiguous. Should the agent execute action p and then observe x, there is no way of knowing whether the agent came from node 1 or node 4. The four columns of $\mathbf{Back}^1$ (fig. 2b) are linearly dependent on $\mathrm{rank}(\mathbf{Back}^1) = 3$. Thus, the rank of $\mathcal{M}^{K,L} = \mathbf{Forw}^K \cdot \mathbf{Back}^L = 3$ for $K \geq 2$ and $L \geq 1$, so we cannot determine the number of states from the SDM rank alone.

Theorem 1 (POMDP Unidentifiability from Trajectories with Reset, Shaw et al. (2026))**.** *There exist two POMDPs* $(b_0, \mathcal{S}, \mathcal{A}, \{T^{a^i} : i \in [n]\}, \{O^{a^i} : i \in [n]\})$ *and* $(\tilde{b}_0, \mathcal{S}, \mathcal{A}, \{\tilde{T}^{a^i} : i \in [n]\}, \{\tilde{O}^{a^i} : i \in [n]\})$ *such that the Markov chain* $\{(s_t, a_t) : a_t \sim \mathrm{Unif}(\mathcal{A}), s_t \sim T^a_{s_{t-1},:}\}_{t=1}^N$ *is ergodic under the distributions of both POMDPs, and for any action sequence* $a_{1:t}$*,*

$$p(o_{1:t} \mid a_{1:t}) = \tilde{p}(o_{1:t} \mid a_{1:t})$$

for all observation sequences $o_{1:t}$, where p and $\tilde{p}$ are the likelihoods of each POMDP, respectively.

Proofs for this claim and all other claims in the main text of this paper can be found in appendix D. When $\mathrm{rank}(\mathcal{M}) < |\mathcal{S}|$, a procedure for finding the true time-homogenous POMDP with a minimal number of states by factoring the SDM $\mathcal{M}$ is not known (Anderson, 1999; Vidyasagar, 2011; Huang et al., 2015). Furthermore, even when $\mathrm{rank}(\mathcal{M}) = |\mathcal{S}|$ and the state space are strongly connected and transitions are aperiodic, according to theorem 1 identifying the POMDP may *still* be impossible. Introducing counterfactuals provides a partial answer, since we assume $\mathbf{Forw}$ must be full-rank, to the first problem and fully resolves the second problem.

3.1 Amplifying SDM Rank with Counterfactuals

To resolve the problem of rank and the number of states, and identifiability as a whole, for the rest of this section, suppose the agent was granted an (admittedly, ridiculous) ability to turn back time. The agent could take actions $a^{(1)}_{t+1:t+L}$ from the unobserved state s_{t+1}, observe $o^{(1)}_{t+1:t+L}$, and then turn time back from $t+L$ to t before the sequence of actions was taken. The agent would take a new set of actions $a^{(2)}_{t+1:t+L}$ and receive observations $o^{(2)}_{t+1:t+L}$ from the same state s_{t+1}, and may even repeat this process to obtain additional action-observation sequences from s_{t+1}. We refer to the collection of tests $\{(ao)^{(1)}_{t+1:t+L}, (ao)^{(2)}_{t+1:t+L}, \dots\}$ from s_{t+1} as a *counterfactual sample*. We will also refer to the set of tests as *counterfactual tests* and use the letter C to denote how many tests are contained within the set, indexing them by parenthesized superscripts $(ao)^{(c)}$, $c \in [C]$. We could estimate an analogous matrix to the SDM above, where the columns are indexed by *sets* of tests that form a possible realization of a counterfactual sample, which we call a *counterfactual system dynamics matrix*. The distribution of each entry is defined as the factorization given by the 'twin-graphical model' of Pearl (2000) applied to a POMDP (see fig. 1b).

Definition 2 (Counterfactual System Dynamics Matrix)**.** *A* Counterfactual *System Dynamics Matrix (CfactSDM) of lengths K and L and number of counterfactuals C for $K, L, C > 0$ is a matrix with $\left(\sum_{i=1}^{K} |\mathcal{A}|^i |\mathcal{O}|^i + 1\right)$ rows and $\left(|\mathcal{A}|^L |\mathcal{O}|^L\right)^C$ columns, whose rows are indexed by* histories *of lengths 0 (null string) to K and columns are indexed by an* ***ordered set*** *of size C of tests of length L. An entry indexed by a history $(ao)_{1:t}$ and a set of tests $\{(ao)^{(c)}_{t+1:t+L}\}_{c=1}^{C}$ is defined as*

$$\mathsf{CfactM}(C)^{K,L}_{(ao)_{1:t},\{(ao)^{(c)}_{t+1:t+L}\}_{c=1}^{C}} = \sum_{s_{t+1}\in\mathcal{S}} p(s_{t+1} \mid (ao)_{1:t}) \prod_{c=1}^{C} p\big(o^{(c)}_{t+1:t+L} \mid s_{t+1}, a^{(c)}_{t+1:t+L}\big). \tag{4}$$

We define $\mathsf{CfactM}(C) = \{\mathsf{CfactM}(C)^{K,L} : K, L > 0\}$, *and* $\mathrm{rank}(\mathsf{CfactM}(C)) = \sup_{K,L>0} \mathrm{rank}(\mathsf{CfactM}(C)^{K,L})$.

Equation (4) has similar structure to the factorization of an SDM entry in eq. (2). When we substitute in the proper expressions for $\mathbf{Forw}$ and $\mathbf{Back}$ matrices, we find that

$$\mathsf{CfactM}(C)^{K,L}_{(ao)_{1:t},\{(ao)^{(c)}_{t+1:t+L}\}_{c=1}^{C}} = \sum_{s_{t+1}\in\mathcal{S}} \mathbf{Forw}^{K}_{(ao)_{1:t},s_{t+1}} \underbrace{\prod_{c=1}^{C} \mathbf{Back}^{L}_{s_{t+1},(ao)^{(c)}_{t+1:t+L}}}_{\mathsf{cfactBack}(C)^L \text{ entry}}. \tag{5}$$

We gather the second term into a new matrix, which we call the *counterfactual back matrix*, or $\mathsf{cfactBack}(C)^L$. Given the relation in eq. (5), vectorizing in the same way as we did for the SDM, we have that $\mathsf{cfactBack}(C)^L = \left(\mathbf{Back}^L\right)^{\odot C}$, where $\odot$ is the *Khatri-Rao product* (elementwise product of all columns, C times over). This product has an intriguing property: given two matrices, A and B with the same number of rows r, $\mathrm{rank}(A \odot B) \geq \min(\mathrm{rank}(A) + \mathrm{rank}(B), r)$ (Bhaskara et al., 2014b). With a sufficient number of counterfactual tests, i.e. $C = O(|\mathcal{S}|)$, and large enough history lengths K and test lengths L, the quantity $\mathrm{rank}(\mathsf{CfactM}(C)^{K,L})$ will be equivalent to the number of POMDP states, as we formalize and show in theorem 2. In practice, one would increase C until the rank of $\mathsf{CfactM}()$ no longer changes to determine the number of states. We can then use the $\mathsf{CfactM}(C)^{K,L}$ to recover the POMDP $\mathbf{Forw}$ and $\mathbf{Back}$ matrices, and the transition and observation matrices after that. To gain intuition for why this works, we return to Merry-Go-Round.

The running example, revisited. The columnwise multiplication acts like a *logical and operation* on the states from which two possible tests could have been emitted. While the four columns of $\mathbf{Back}^1$ are not full rank, if we take the elementwise products of columns, we find that the 4×4 identity matrix to a submatrix of $\mathsf{cfactBack}(2)^1 = \mathbf{Back}^1 \odot \mathbf{Back}^1$ (fig. 2c), implying the matrix is full-rank. Since $\mathbf{Forw}^K$ of Merry-Go-Round is full-rank for $K \geq 1$, then $\mathrm{rank}(\mathsf{CfactM}(2)^{K,L}) = 4$ for $K \geq 1, L \geq 1$, the correct number of states of the POMDP.

We now have a strategy for learning general POMDPs, provided $\mathbf{Forw}$ is full-rank and the agent has the ability to turn back time. The agent, under uniform random exploration would: execute a random action trajectory of length K to form a *history*, execute a random action trajectory of length L to gather one *counterfactual test*, turn back time L steps, gather another counterfactual test, and repeat $C-2$ more times for the rest of the counterfactual sample. The agent would then execute another set of actions of length K and continue. In the limit, we could estimate $\mathsf{CfactM}(C)^{K,L}$ exactly by frequency counting co-occurences of histories and counterfactual samples. In this setting, we can rely on *tensor decomposition methods* to identify the POMDP.

3.2 POMDP Recovery with Tensor Decomposition

First, we reshape $\mathsf{CfactM}(C)^{K,L}$ from a two-dimensional *matrix* that is row-indexed by histories and column-indexed by C-tuples of counterfactual tests into a $(C+1)$-mode *tensor* T. The first mode of this tensor is indexed by histories, and the remaining C modes are indexed by counterfactual tests:

$$\mathsf{T}_{(ao)_{1:t},(ao)^{(1)}_{t+1:t+L},(ao)^{(2)}_{t+1:t+L},\ldots,(ao)^{(C)}_{t+1:t+L}} = \mathsf{CfactM}(C)^{K,L}_{(ao)_{1:t},\{(ao)^{(c)}_{t+1:t+L}\}_{c=1}^{C}}$$

We then set $C = 2|\mathcal{S}|$, which is far more counterfactual tests than usually required but is *sufficient* to ensure POMDP recovery by tensor decomposition methods. This decomposition reveals $\mathbf{Forw}^K$, and $\mathsf{cfactBack}(|\mathcal{S}|)^L$ which can be used to recover the original POMDP.

Theorem 2 (POMDP Recoverability from Counterfactuals)**.** *Given as input* $\mathsf{CfactM}(C)^{K,L}$*, the tensor decomposition algorithm of Anandkumar et al. (2012) will return transition matrices* $\{T^{a^i} : i \in [n]\}$ *and observation matrices* $\{O^{a^i} : i \in [n]\}$ *up to a permutation of states.*

Obtaining counterfactual samples, then is *sufficient* (theorem 2) to identify minimal time-homogeneous POMDPs with full-rank $\mathbf{Forw}$ matrices from actions and observations. This result resolves questions of identifiability conditions for a broader class of time-homogenous POMDPs than discussed in prior work (Vidyasagar, 2011; Huang et al., 2015).

4 Robustly Learning Deterministic POMDPs

Of course, agents that observe a single trajectory *cannot* arbitrarily go back in time to collect counterfactual data. However, *deterministic* POMDPs (essentially, deterministic finite state automata) do exhibit natural reset mechanisms. Under the assumption there exists a sequence of actions that direct any state to any other state in the POMDP, there exists *anchor sequences*: action-observation histories that determine the state when observed (fig. 2b, lemma 4). The agent, of course, has no knowledge of what the anchor sequences are at the start of the learning problem. Our process for discovering these sequences requires ideas from nonnegative matrix factorization (Arora et al., 2013; Gillis, 2020) and the rank-amplification properties of counterfactuals, which will culminate in our algorithm, Counterfactual Successive Projection (CfactSP). For section 4.1, we will assume there is no noise and that the SDM we have is *exact*. We relax these assumptions in section 4.2.

4.1 Discovering Anchor Sequences

Our first step is to reinterpret the factorization of the SDM $\mathcal{M}^{K,L}$ into $\mathbf{Forw}^K$ and $\mathbf{Back}^L$ matrices in a geometric sense. First, we observe that from our derivation of the nonnegative matrix factorization that every row of the SDM is a convex combination of the rows of $\mathbf{Back}^L$ with convex weights encoded as rows of $\mathbf{Forw}^K$. Next, due to the existence of anchor sequences, we observe that the rows of $\mathbf{Back}^L$ themselves form a subset of the rows of the SDM (fig. 2d). Furthermore, anchor sequences imply that a copy of the identity matrix is a submatrix of $\mathbf{Forw}^K$, so $\mathbf{Forw}^K$ is full-rank

(fig. 2b). If we reinterpret the rows of the SDM as high-dimensional vectors, then the vertices of the *convex hull* of the rows of the SDM form the $\mathbf{Back}^L$ matrix, and the histories that index those vertices are anchor sequences. To compute the convex hull, we rely on a class of methods for non-negative matrix factorization (Arora et al., 2013), called *successive projection algorithms* (Gillis, 2020). These methods iteratively search for the row of $\mathcal{M}^{K,L}$ with maximal norm then project all rows to the orthogonal complement of the row to eliminate near-duplicates, and stop when the projected norm of the remaining vectors is sufficiently small.

Successive projection methods, often assume the vertices of the convex hull are linearly independent, which is not guaranteed for $\mathbf{Back}^L$. (Arora et al., 2013). Merry-Go-Round, for example, is a deterministic POMDP with a rank-deficient $\mathbf{Back}^L$ matrix, even when expressed in minimal form. To address this problem, we again rely on counterfactuals. While we still cannot compute the true counterfactual distributions to form a $\mathsf{CfactM}(C)^{K,L}$ matrix *for all* histories, we can exploit the fact that the existence of anchor sequences ensures that $\mathbf{Back}^L$ is a submatrix of $\mathcal{M}^{K,L}$. If we take Khatri-Rao products of the $\mathcal{M}^{K,L}$ for a given number of counterfactual tests C, then we will, implicitly, take Khatri-Rao products of the submatrix $\mathbf{Back}^L$ and amplify its rank.

Our algorithm, Counterfactual Successive Projection, outlined in algorithm 1, combines the ideas above. The algorithm first computes Khatri-Rao products of the SDM, runs successive projection, and then checks to see if all the rows of the SDM lie within the found convex hull as a stopping condition. For ease of exposition, algorithm 1 only handles for the noiseless case with an exact SDM. Next, we relax both assumptions.

Algorithm 1 Counterfactual Successive Projection (CfactSP)

Require: Trajectory $\mathcal{D} = (ao)_{1:N}$; maximum history length K; test length L; number of counterfactuals C.

$\hat{\mathcal{M}}^{K,L} \leftarrow$ EstimateSDM$(\mathcal{D}, K, L)$ using eq. (6)
$\mathcal{H}_{\text{cand}} \leftarrow \bigcup_{t=0}^{K-1} (\mathcal{A} \times \mathcal{O})^t$
$X \leftarrow (\hat{\mathcal{M}}^{K,L}_{\mathcal{H}_{\text{cand}},:})^{\odot C}$ ▷ Compute counterfactuals.
$R \leftarrow X$, $\mathcal{H}_{\text{sel}} \leftarrow ()$, and $\mathcal{H}_{\text{active}} \leftarrow \mathcal{H}_{\text{cand}}$
while $\mathcal{H}_{\text{active}} \neq \emptyset$ **do**
  $h^\star \leftarrow \arg\max_{h \in \mathcal{H}_{\text{active}}} \|R_{:,h}\|_2$ ▷ Pick row with maximum norm.
  $\mathcal{H}_{\text{active}} \leftarrow \mathcal{H}_{\text{active}} \setminus \{h^\star\}$
  $u \leftarrow R_{:,h^\star}$ and $R \leftarrow \left(I - uu^\top / \|u\|_2^2\right) R$ ▷ Project other rows.
  append $h^\star$ to $\mathcal{H}_{\text{sel}}$ ▷ Stop if all rows are inside convex hull of $\mathcal{H}_{\text{sel}}$.
  if $\max_{h \in \mathcal{H}_{\text{cand}}} \operatorname{dist}\left(\hat{\mathcal{M}}^{K,L}_{h,:}, \operatorname{conv}\{\hat{\mathcal{M}}^{K,L}_{h',:} : h' \in \mathcal{H}_{\text{sel}}\}\right) = 0$ **then**
    break
$\hat{\mathcal{P}} \leftarrow$ ReconstructPOMDP$(\mathcal{H}_{\text{sel}}, \hat{\mathcal{M}}^{K,L})$

Ensure: $\hat{\mathcal{P}}$

4.2 Handling Finite Samples and Random Corruptions

Deterministic environments, of course, do not model real-world dynamics well. Instead, we assume that the environments are *near-deterministic* with a small amount of random noise.

Definition 3 (ϵ-corrupted POMDP)**.** *Let $\mathcal{P} = (b_0, \{T^a : a \in \mathcal{A}\}, \{O^a : a \in \mathcal{A}\})$ be a deterministic POMDP, i.e., every entry of T^a and O^a is 0 or 1 for all $a \in \mathcal{A}$. For $\epsilon > 0$, a POMDP $\tilde{\mathcal{P}} = (\tilde{b}_0, \{\tilde{T}^a : a \in \mathcal{A}\}, \{\tilde{O}^a : a \in \mathcal{A}\})$ over the same states, actions, and observations is an* ϵ-corrupted POMDP *of $\mathcal{P}$ if*

$$\max_{a \in \mathcal{A},\, i \in [r]} \left\| T^a_{i,:} - \tilde{T}^a_{i,:} \right\|_{\text{TV}} \leq \epsilon \qquad \textit{and} \qquad \max_{a \in \mathcal{A},\, i \in [r]} \left\| O^a_{i,:} - \tilde{O}^a_{i,:} \right\|_{\text{TV}} \leq \epsilon.$$

We want to recover the original deterministic POMDP from a trajectory gathered from an ϵ-corrupted POMDP up to a permutation of states.

Given a finite data trajectory $\mathcal{D} = (ao)_{1:N}$ (where $a_i \sim \text{Unif}(\mathcal{A})$ i.i.d. $\forall i \in [N]$), we can estimate the entries of an SDM $\hat{\mathcal{M}}^{K,L}$ by frequency counting over sliding windows of the trajectory (using

the suffix-counting estimator of Wolfe et al. (2005)):

$$\hat{\mathcal{M}}^{K,L}_{(a'o')_{1:t},(a'o')_{t+1:t+L}} = \frac{\sum_{i=0}^{N-(t+L)} \mathbb{I}_{(a'o')_{1:t+L}=(ao)_{i+1:i+t+L}}}{\sum_{i=0}^{N-(t+L)} \mathbb{I}_{(a'o')_{1:t}=(ao)_{i+1:i+t}} \cdot \mathbb{I}_{a'_{t+1:t+L}=a_{i+t+1:i+t+L}}}. \tag{6}$$

The introduction of noise and finite samples introduces two main concerns. The first is that there is a small probability that an observed history may be corrupted to produce a string that would never be yielded by the original deterministic POMDP. The appearances of these histories are quite rare and will consist of entries that have only been incremented once by data, resulting in many entries of 1s and 0s. These rows can have larger norm relative to true anchor histories and get picked by the successive projection step of algorithm 1, resulting in an incorrect output. The second is that the **Back** matrix of the ϵ-corrupted POMDP is slightly different from that of the deterministic POMDP and must be post-processed in some way.

Remarkably, algorithm 1 can be made robust to a small amount of corruption with a few changes. First, we *filter out* a candidate history unless every action-sequence block has at least $\nu \cdot N$ visits, where N is the number of steps taken by the agent and ν is some frequency threshold. When set appropriately, this threshold eliminates erroneous rare rows while preserving the rows needed for recovery. Second, after choosing the row of maximal norm, we round each of its entries to the nearest deterministic row for the projection and hull test, as well as when reconstructing the POMDP. Putting these steps together, we get the following robustness guarantee, with bounds using additional POMDP-dependent quantities proved in section D.8.

Theorem 3 (CfactSP, informal)**.** *Let $\mathcal{P}$ be a deterministic POMDP with r states, n actions, and m observations, in which every state is reachable from every other state and every pair of distinct states produces different observation sequences under some common finite action sequence. Suppose the data trajectory $\mathcal{D} = (ao)_{1:N}$ is collected using independently and uniformly sampled actions from an ϵ-corrupted POMDP $\tilde{\mathcal{P}}$ of $\mathcal{P}$ (definition 3). There are universal constants $c_1, c_2 > 0$ such that, for any $\delta \in (0, 1)$, there are choices of the history and test lengths K, L, the number of counterfactual tests C, and the frequency-filter and stopping parameters for which*

$$\epsilon \leq \max\{2, n\}^{-c_1 r^2}, \qquad N \geq \max\{2, nm\}^{c_2 r^2} \log \frac{2}{\delta}$$

ensure that the robust CfactSP procedure in section D.1 returns exactly r states and the uncorrupted transition and observation matrices up to state permutation, with probability at least $1 - \delta$.

The derivation in section D.8.3 uses $L = \max\{1, r - 1\}$ and a history cap $K \leq r^2$. The maximum with two keeps these sufficient bounds meaningful when there is only one action.

5 Numerical Simulations

To evaluate the viability of CfactSP in practice, we attempt to learn POMDPs from trajectories generated from synthetic domains under action-substitution noise (`act-sub`). At each timestep, with probability $\epsilon > 0$, we replace the action chosen by a uniform random exploration policy with another action drawn uniformly from $\mathcal{A}$. The agent still records the original action chosen by the exploration policy. Additional corruption models are described in appendix B.

In fig. 3, we show recovery rates for a domain called *toolbox*, a partially observable domain with five states associated with a drawer and lid that can be open or closed with a locking mechanism, two actions to toggle the drawer or lid, and four observations that describe whether the drawer and lid are open or closed. The deterministic locking mechanism can fail under noise corruption. We see that in 2.0×10^4 actions we can recover the structure under mild noise levels. In comparison to our tightest sample complexity bound theorem 9, for an *uncorrupted* toolbox it gives approximately 2.1×10^{10} samples at a 95% recovery rate, a gap of roughly six orders of magnitude. Descriptions of additional domains and corruption models appear in appendices A and B.

6 Related Work

Learning algorithms for deterministic POMDPs have been widely studied, especially in the noiseless case, and often have identifiability guarantees. A seminal result in learning deterministic automata

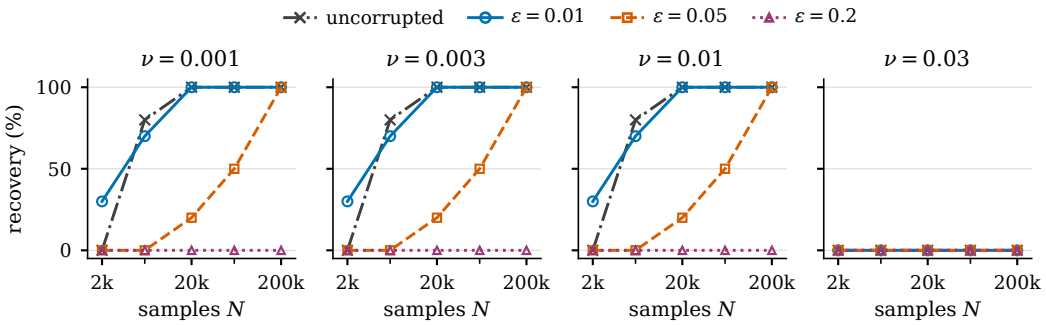


Figure 3: Recovery of the toolbox model under action corruption ϵ and rare event filtering rates ν, measured over 10 seeds. The recovery curves are identical for $\nu = 0.001$, 0.003, and 0.01; no run recovers at $\nu = 0.03 > \nu^* \approx 0.01713$, the frequency of the rarest anchor sequence.

is the L* algorithm (Angluin, 1987), which assumes a reset to a start state. An algorithm of Rivest & Schapire (1993) relaxes this assumption by leveraging homing experiments (Kohavi & Jha, 2009) to invoke a reset, but is also sensitive to corrupted data. Dean et al. (1995) develop an approach that identifies a deterministic model from corrupting noise but also requires knowledge of the POMDP's *distinguishing sequences*, which are tests that identify the initial state of the POMDP before the test was observed. Not every POMDP, however, has distinguishing sequences (Kohavi & Jha, 2009).

To learn predictive models of POMDPs, a common approach is to take a spectral decomposition of the SDM. These methods yield models that predict the likelihood of a future trajectory in the limit of infinite data (Jaeger, 1998; Littman & Sutton, 2001; Boots et al., 2011) and often have polynomial sample complexities in some cases, such as for Hidden Markov Models (Hsu et al., 2012; Balle et al., 2014). Given technical conditions on the rank of transition or observation matrices, tensor decomposition methods can be leveraged to identify the underlying POMDP, with polynomial sample and runtime guarantees (Azizzadenesheli et al., 2016; Guo et al., 2016). We provide an intuitive interpretation for the rank conditions in this paper using the language of counterfactuals (Pearl, 2000; Bhaskara et al., 2014a).

Recent approaches use deep learning to learn *world models*, or latent representations of an environment trained on action-observation sequences, where observations are images (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2025; Agarwal et al., 2025; Wu et al., 2022; Zhou & Zhang, 2026). Aside from the work of Zhang et al. (2026), which provides identifiability guarantees for world models assuming an invertible state-to-observation map, these studies are largely empirical and are focused on learning a latent representation useful for planning.

7 Conclusion

We have studied the problem of identifying a discrete time-homogenous POMDP from a single trajectory of actions and observations. We showed that *counterfactual* samples, or joint distributions over observations over multiple courses of action executed from the same state, are necessary and sufficient to identify POMDPs with full-rank **Forw** matrices. In general POMDPs, these counterfactual samples are impossible to form, but deterministic POMDPs offer a natural reset mechanism that allows for the construction of counterfactual samples from a single trajectory. We exploit this reset mechanism in our algorithm, CfactSP, which we show to learn deterministic POMDPs despite low-likelihood random corruptions to state transitions and observation emissions. Our empirical results suggest our algorithm is more sample efficient than our theoretical results state.

In the future, there are a number of directions to improve our work and yield an algorithm applicable to real decision-making agents. First, our sample complexity guarantees are rather high (exponentially complex in POMDP constants). A fruitful future line of work is to investigate *nonuniform* exploration strategies that may potentially yield better bounds and practical performance (Bowling et al., 2006). Furthermore, while we show that *learning* a deterministic model is possible under noise, we must also allow agents to leverage these models to *plan*, even under corruptions.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing proofs, providing critical ingredients for proving mathematical claims, and implementing methods in code for numerical simulations. We have not used generative AI tools to generate synthetic data sets, develop theoretical models or conceptual frameworks, design or provide feedback on research methodology, experiments, data analysis, or interpreting results. Use of LLMs to assist with translation is not applicable to this work. Additionally, we used generative AI tools to create the pictograms in Figure 4 (ChatGPT Images 2.5), describe synthetic domains in appendix A, and generate TikZ code for Figure 2a. We have reviewed all AI-assisted work. All mathematical proofs have been checked by hand separately, and any cited work used in a proof has been independently verified for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We do not believe this work to be cause of any ethical concern.

REPRODUCIBILITY STATEMENT

All omitted proofs to derive the theoretical claims in this paper have been included in appendices. Code, including synthetic domains and the learning algorithm, will be made available should our work be accepted for publication.

REFERENCES

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Cameron Allen, Aaron Kirtland, Ruo Yu Tao, Sam Lobel, Daniel Scott, Nicholas Petrocelli, Omer Gottesman, Ronald Parr, Michael Littman, and George Konidaris. Mitigating Partial Observability in Sequential Decision Processes via the Lambda Discrepancy. *Advances in Neural Information Processing Systems*, 37:62988–63028, December 2024.

Anima Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor Decompositions for Learning Latent Variable Models:. Technical report, Defense Technical Information Center, Fort Belvoir, VA, December 2012.

Brian D. O. Anderson. The Realization Problem for Hidden Markov Models. *Mathematics of Control, Signals and Systems*, 12(1):80–120, April 1999. ISSN 1435-568X. doi: 10.1007/PL00009846.

Dana Angluin. Learning regular sets from queries and counterexamples. *Information and Computation*, 75(2):87–106, November 1987. ISSN 0890-5401. doi: 10.1016/0890-5401(87)90052-6.

Sanjeev Arora, Rong Ge, Yonatan Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. A Practical Algorithm for Topic Modeling with Provable Guarantees. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 280–288. PMLR, May 2013.

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement Learning of POMDPs using Spectral Methods. In *Conference on Learning Theory*, pp. 193–256. PMLR, June 2016.

Borja Balle, Xavier Carreras, Franco M. Luque, and Ariadna Quattoni. Spectral learning of weighted automata. *Machine Learning*, 96(1):33–63, July 2014. ISSN 1573-0565. doi: 10.1007/s10994-013-5416-x.

Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '14, pp. 594–603, New York, NY, USA, May 2014a. Association for Computing Machinery. ISBN 978-1-4503-2710-7. doi: 10.1145/2591796.2591881.

Aditya Bhaskara, Moses Charikar, and Aravindan Vijayaraghavan. Uniqueness of Tensor Decompositions with Applications to Polynomial Identifiability. In *Proceedings of The 27th Conference on Learning Theory*, pp. 742–778. PMLR, May 2014b.

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, June 2011. ISSN 0278-3649. doi: 10.1177/0278364911404092.

Michael Bowling, Peter McCracken, Michael James, James Neufeld, and Dana Wilkinson. Learning predictive state representations using non-blind policies. In *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, pp. 129–136, Pittsburgh, Pennsylvania, 2006. ACM Press. ISBN 978-1-59593-383-6. doi: 10.1145/1143844.1143861.

Thomas Dean, Dana Angluin, Kenneth Basye, Sean Engelson, Leslie Kaelbling, Evangelos Kokkevis, and Oded Maron. Inferring finite automata with stochastic output functions and an application to map learning. *Machine Learning*, 18(1):81–108, January 1995. ISSN 1573-0565. doi: 10.1007/BF00993822.

Nicolas Gillis. *Nonnegative Matrix Factorization*. Data Science. Society for Industrial and Applied Mathematics, January 2020. ISBN 978-1-61197-640-3. doi: 10.1137/1.9781611976410.

Zhaohan Daniel Guo, Shayan Doroudi, and Emma Brunskill. A PAC RL Algorithm for Episodic POMDPs. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 510–518. PMLR, May 2016.

David Ha and Jürgen Schmidhuber. Recurrent World Models Facilitate Policy Evolution. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *International conference on machine learning*, pp. 2555–2565. PMLR, 2019.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse control tasks through world models. *Nature*, 640(8059):647–653, 2025.

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, September 2012. ISSN 0022-0000. doi: 10.1016/j.jcss.2011.12.025.

Qingqing Huang, Rong Ge, Sham Kakade, and Munther Dahleh. Minimal Realization Problems for Hidden Markov Models, December 2015.

Masoumeh T. Izadi and Doina Precup. Point-Based Planning for Predictive State Representations. In Sabine Bergler (ed.), *Advances in Artificial Intelligence*, pp. 126–137, Berlin, Heidelberg, 2008. Springer. ISBN 978-3-540-68825-9. doi: 10.1007/978-3-540-68825-9_13.

Herbert Jaeger. Discrete-time, Discrete-valued Observable Operator Models: A Tutorial. Technical Report, European Research Consortium for Informatics and Mathematics at FHG, October 1998.

Zvi Kohavi and Niraj K. Jha. *Switching and Finite Automata Theory*. Cambridge University Press, Cambridge, 3 edition, 2009. ISBN 978-0-521-85748-2. doi: 10.1017/CBO9780511816239.

Michael Littman and Richard S Sutton. Predictive Representations of State. In *Advances in Neural Information Processing Systems*, volume 14. MIT Press, 2001.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, Cambridge, U.K. ; New York, 2000. ISBN 978-0-521-89560-6 978-0-521-77362-1.

R. L. Rivest and R. E. Schapire. Inference of Finite Automata Using Homing Sequences. *Information and Computation*, 103(2):299–347, April 1993. ISSN 0890-5401. doi: 10.1006/inco.1993.1021.

Seiji Shaw, Travis Manderson, Chad Kessens, and Nicholas Roy. Toward Learning POMDPs Beyond Full-Rank Actions and State Observability, February 2026.

Satinder Singh, Michael James, and Matthew Rudary. Predictive State Representations: A New Theory for Modeling Dynamical Systems, July 2012.

Joel A. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011. URL `https://tropp.caltech.edu/papers/Tro11-Freedmans-Inequality.pdf`.

M. Vidyasagar. The complete realization problem for hidden Markov models: A survey and some new results. *Mathematics of Control, Signals, and Systems*, 23(1):1–65, December 2011. ISSN 1435-568X. doi: 10.1007/s00498-011-0066-7.

Britton Wolfe, Michael R. James, and Satinder Singh. Learning predictive state representations in dynamical systems without reset. In *Proceedings of the 22nd International Conference on Machine Learning - ICML '05*, pp. 980–987, Bonn, Germany, 2005. ACM Press. ISBN 978-1-59593-180-1. doi: 10.1145/1102351.1102475.

Philipp Wu, Alejandro Escontrela, Danijar Hafner, Ken Goldberg, and Pieter Abbeel. Daydreamer: World models for physical robot learning. *Conference on Robot Learning*, 2022.

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.

Todd Y Zhou and Daniel Zhang. Learning counterfactual world models for embodied reasoning under partial observability. *arXiv preprint arXiv:2609.05834*, 2026.

A Appendix: Synthetic Domains

We evaluate on four finite deterministic POMDPs: Toolbox, Espresso, Washroom, and Merry-Go-Round. The first three are inspired by household appliances. Below we specify their complete state, action, and observation spaces and describe how their dynamics are encoded. Figures 2a and 4 show the corresponding transition diagrams.

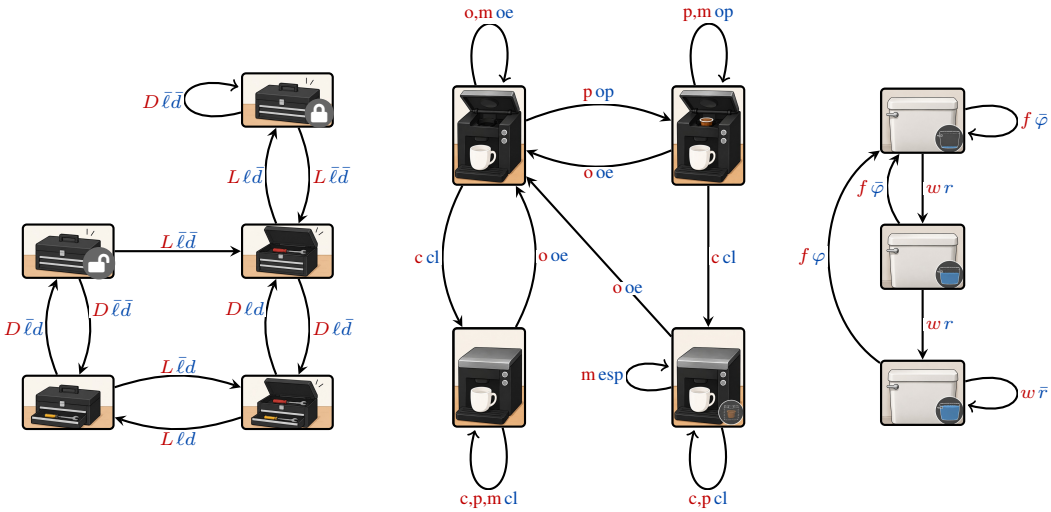
Model representation and trajectory generation. Each model is specified by a next-state map $f_a : \mathcal{S} \to \mathcal{S}$ and an emitted-observation map $g_a : \mathcal{S} \to \mathcal{O}$ for each action a. Every action is available in every state. In the notation of section 2, these give

$$T^a_{ij} = \mathbf{1}\{f_a(s^i) = s^j\}, \qquad O^a_{io} = \mathbf{1}\{g_a(s^i) = o\}.$$

The implementation stores an $|\mathcal{S}| \times |\mathcal{S}|$ transition matrix for each action and diagonal emission matrices $D^{a,o}$ for each action–observation pair, whose diagonal entries indicate which states emit o under a. Toolbox and Washroom index these emission matrices by the state being left, whereas Espresso and Merry-Go-Round index them by the state reached. The joint matrix for emitting o and moving to the next state is

$$J^{a,o}_{ij} = \mathrm{P}(o_t = o, s_{t+1} = s^j \mid s_t = s^i, a_t = a),$$
$$J^{a,o} = \begin{cases} D^{a,o}T^a & \text{for emissions from the state being left,} \\ T^a D^{a,o} & \text{for emissions from the state reached.} \end{cases}$$

For the latter convention, if $\widetilde{g}_a$ is the reading at the destination, we set $g_a(s) = \widetilde{g}_a(f_a(s))$ to obtain the pre-transition observation model used in section 2. To generate trajectories from these models, we sample actions independently and uniformly. Each trajectory begins with an initial state sampled from b_0, the stationary distribution under uniform random actions, and then proceeds without resets.



(a) **Toolbox.** Actions toggle the lid (L) or the drawer (D). The observation is the lid/drawer configuration of the state being *left*, with ℓ for lid open, d for drawer open, and a bar for closed. The lock (badge) is never directly observed as the two closed states look identical. Closing the lid over a closed drawer locks it.

(b) **Espresso.** Actions are open, close, put-pod, make. The observation is the state *arrived* in: oe open/empty, op open/pod, cl closed (pod hidden), esp closed with espresso dispensed. The two closed states look identical from outside. Opening discards the pod.

(c) **Washroom** ($F = 2$). wait advances the refill and reports whether water is running. flush always empties the tank and reports φ only if it was full.

Figure 4: **Synthetic domains.** State transition diagrams for three of the synthetic domains we evaluate on in this work. Edge labels read action observation as in fig. 2a.

A.1 TOOLBOX

A state is a triple (ℓ, d, k) indicating whether the lid is open, the drawer is open, and the drawer is locked, respectively, with 1 denoting true. The five possible states are

$$\mathcal{S} = \{(0,0,1),\ (1,0,0),\ (1,1,0),\ (0,1,0),\ (0,0,0)\}.$$

The remaining three combinations are excluded as the drawer can be locked only when both the lid and the drawer are closed. The two actions are $\mathcal{A} = \{\texttt{toggle_lid}, \texttt{toggle_drawer}\}$ (respectively L, D in fig. 4a). Opening the lid unlocks the drawer and closing the lid locks the drawer if the drawer is closed. Toggling a locked drawer leaves the state unchanged. The four possible observations are $\mathcal{O} = \{(0,0),(1,0),(1,1),(0,1)\}$, and $g_a(\ell, d, k) = (\ell, d)$ for either action: the observation reports the lid and drawer configuration *before* the action, and never reports the lock. Toolbox provides a relatively approachable test of recovering a model of a system with a hidden lock: observations are independent of the chosen action, and every state has an anchor sequence of just one action–observation pair.

A.2 ESPRESSO

A state is a pair (c, ℓ) indicating whether a coffee pod is present and whether the lid is open. There are four states and four actions:

$$\mathcal{S} = \{(1,0),\ (1,1),\ (0,0),\ (0,1)\}, \qquad \mathcal{A} = \{\texttt{open}, \texttt{close}, \texttt{put-pod}, \texttt{make}\}.$$

The `open` action opens the lid and removes any pod, including when the lid is already open. The `close` action closes the lid and preserves the pod. The `put-pod` action inserts a pod if the lid is open and otherwise leaves the state unchanged. The `make` action leaves the state unchanged and produces espresso exactly when the lid is closed and a pod is present. Making espresso does not consume the pod in this model.

An observation is a triple (visible pod status, lid status, espresso output). The complete alphabet used by the experiments is

$$\mathcal{O} = \big\{(\text{out}, \text{open}, 0),\ (\text{in}, \text{open}, 0),\ (\text{na}, \text{closed}, 0),\ (\text{na}, \text{closed}, 1)\big\}.$$

Here na means that pod presence is hidden by the closed lid, and the final bit indicates whether espresso was dispensed by the action. These four symbols are labeled *oe*, *op*, *cl*, and *esp*, respectively, in fig. 4b. Lid and visible pod status are read from the state *after* the action; the espresso bit is one only for a successful `make` action. Espresso is challenging because the sequence of events leading to the $(\texttt{make}, esp)$ anchor is rare during random exploration, making the anchor difficult to distinguish from transition or observation noise.

A.3 WASHROOM

Washroom tests recovery when some states require anchor sequences longer than one action–observation pair. The hidden state is the tank's fill level $i \in \mathcal{S} = \{0, 1, \ldots, F\}$, where 0 is empty and F is full. We use $F = 2$, giving three states: empty, partially full, and full. The two actions and four observations are

$$\mathcal{A} = \{\texttt{flush}, \texttt{wait}\}, \qquad \mathcal{O} = \{\texttt{flush}, \texttt{no-flush}, \texttt{running}, \texttt{no-running}\}.$$

Flushing always empties the tank, $f_{\texttt{flush}}(i) = 0$, and emits `flush` if $i = F$ and `no-flush` otherwise. Waiting advances the fill level, $f_{\texttt{wait}}(i) = \min(i + 1, F)$, and emits `running` if $i < F$ and `no-running` otherwise. These observations are indexed by the state *before* the action. The fill level is not directly observed, although a `no-running` observation reveals that the tank is full. For $F = 2$, a flush followed by a wait anchors the partially full state, while two consecutive waits guarantee that the tank is full.

A.4 MERRY-GO-ROUND

Using the labels in fig. 2a, the domain has four states, two actions, and two observations: $\mathcal{S} = \{1, 2, 3, 4\}$, $\mathcal{A} = \{p, q\}$, and $\mathcal{O} = \{x, y\}$. The complete transition dynamics are shown in fig. 2a. The observation reports the state reached: destinations $1, 2$ emit x, and destinations $3, 4$ emit y, independently of the action given the destination. Each action–observation pair uniquely identifies the destination state, although it does not uniquely identify the state being left.

B Additional Recovery Results

In addition to the `act-sub` model described in section 5, we consider two variants that corrupt only observation generation or state transitions. At each timestep, with probability ϵ, we draw a substitute action uniformly from $\mathcal{A}$. Under `obs-spoof`, observation generation uses the substitute action, while the state transition uses the agent's chosen action. Under `trans-spoof`, the state transition uses the substitute action, while observation generation uses the agent's chosen action. In both variants, the agent records the original action chosen by the exploration policy.

Figure 5 gives the `act-sub` frequency-threshold sweep for washroom, toolbox, espresso, and merry-go-round, including the toolbox measurements shown in fig. 3. Each point reports the percentage of 10 seeds that recover the uncorrupted POMDP. We use sample sizes $N \in \{2000, 6000, 20000, 60000, 200000\}$ and frequency thresholds $\nu \in \{0.001, 0.003, 0.01, 0.03\}$. For each setting, we plot uncorrupted trajectories and `act-sub` corruption at $\epsilon \in \{0.01, 0.05, 0.20\}$. All panels use the same algorithm settings as the main figure.

The reference frequency ν^* shown for each domain is the minimum uncorrupted visit frequency over the designated anchor histories and test-action groups, matching the per-group frequency filter. These reference values are approximately 0.03061 for washroom, 0.01713 for toolbox, 0.01026 for espresso, and 0.1242 for merry-go-round. The threshold $\nu = 0.03$ exceeds the toolbox and espresso reference frequencies, and neither domain recovers at that threshold in these experiments. Espresso at $\nu = 0.01$ is close to its reference frequency and exhibits variable recovery even on uncorrupted trajectories, so insensitivity to the filter rate should not be expected near this boundary. Merry-go-round recovers in every plotted setting, with all thresholds well below its reference frequency.

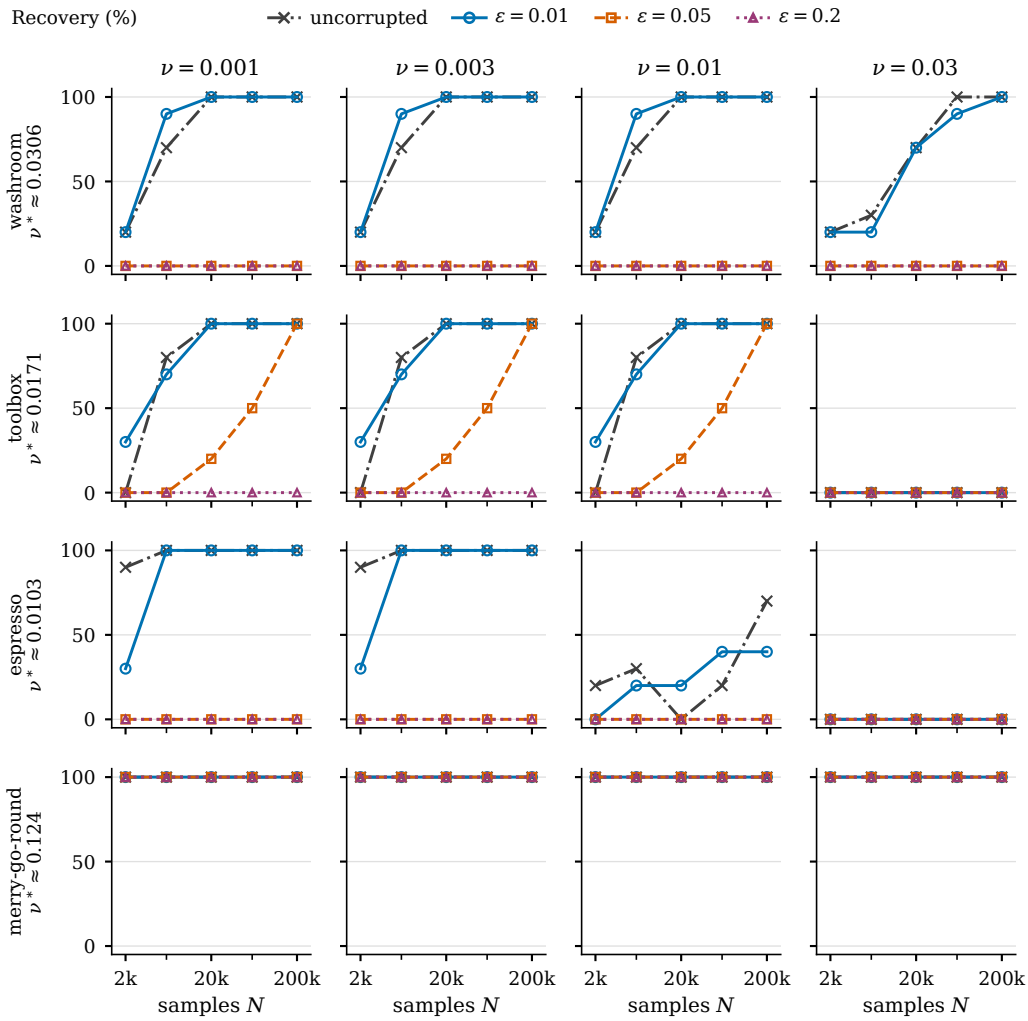


Figure 5: Recovery under `act-sub` corruption for washroom, toolbox, espresso, and merry-go-round. Rows vary the domain, columns vary the frequency filter rate ν, and curves vary the corruption rate ϵ, including an uncorrupted baseline. Each point is the percentage of successful recoveries over 10 seeds. The reference frequency ν^* of the rarest anchor, measured per test-action group, is shown beside each domain.

C Proof of Counterfactual Recoverability

This section gives a precise statement and proof of the recovery guarantee in theorem 2 under the full **Forw** rank assumption. Forward and counterfactual SDM matrices below are restricted to histories with positive probability under their prescribed action sequences.

Lemma 1 (Counterfactual tensor decomposition)**.** *Suppose* $\mathbf{Back}^L$ *has pairwise distinct rows,* $\operatorname{rank}(\mathbf{Forw}^K) = r$*, and* $C \geq 2r$*, and set* $q = \lfloor C/2 \rfloor$*. Let* T *be the three-mode tensor obtained from* $\mathsf{CfactM}(2q)^{K,L}$ *by grouping its test coordinates into two groups of* q *tests, with histories indexing the first mode. Then*

$$\mathsf{T} = \sum_{s=1}^{r} \mathbf{Forw}^K_{:,s} \otimes (\mathsf{cfactBack}(q)^L_{s,:})^\top \otimes (\mathsf{cfactBack}(q)^L_{s,:})^\top, \tag{7}$$

and the three factor matrices $\mathbf{Forw}^K, (\mathsf{cfactBack}(q)^L)^\top, (\mathsf{cfactBack}(q)^L)^\top$ *have full column rank* r*. Moreover,* $\operatorname{rank}(\mathsf{CfactM}(C)^{K,L}) = r$*.*

Proof. If C is odd, obtain $\mathsf{CfactM}(2q)^{K,L}$ from the supplied matrix by fixing the unused test's action sequence and summing over its observation sequences. For even C, all tests are retained. The factorization in (5), with the tests divided into two groups, gives (7).

The column Kruskal rank of a matrix is the largest integer k, at most its number of columns, for which every set of k columns is linearly independent. For two matrices with r nonzero columns and column Kruskal ranks k_1, k_2, their column-wise Khatri-Rao product has Kruskal rank at least $\min\{k_1 + k_2 - 1, r\}$ (Bhaskara et al., 2014a, Lemma 2.5). Apply this bound to the transposes of the paper's row-wise Khatri-Rao products.

Each row of $\mathbf{Back}^L$ has sum n^L, since each action-sequence block sums to one. Distinct rows therefore cannot be proportional, so $\operatorname{krank}((\mathbf{Back}^L)^\top) \geq 2$ when $r \geq 2$. Repeated application of the bound gives, for every $\ell \geq 1$,

$$\begin{aligned}\operatorname{krank}\big((\mathsf{cfactBack}(\ell)^L)^\top\big) &\geq \min\{\ell(\operatorname{krank}((\mathbf{Back}^L)^\top) - 1) + 1, r\} \\ &\geq \min\{\ell + 1, r\}.\end{aligned}$$

Since $q \geq r$, the matrix $\mathsf{cfactBack}(q)^L$ has full row rank r. For $r = 1$, it has a single nonzero row, so the conclusion holds as well. The history factor has full column rank by assumption. The same argument makes $\mathsf{cfactBack}(C)^L$ full row rank, so

$$\operatorname{rank}(\mathsf{CfactM}(C)^{K,L}) = \operatorname{rank}\big(\mathbf{Forw}^K \mathsf{cfactBack}(C)^L\big) = r.$$

□

Theorem 4 (Precise counterfactual recovery guarantee)**.** *Suppose* $\mathbf{Back}^L$ *has pairwise distinct rows,*

$$\operatorname{rank}(\mathbf{Forw}^{K-1}) = r, \qquad C \geq 2r.$$

Then the exact $\mathsf{CfactM}(C)^{K,L}$ *determines* r*,* $\mathbf{Back}^L$*,* $\mathbf{Forw}^K$*, and all transition and observation matrices* T^a, O^a*, up to one common permutation of the states. A tensor decomposition followed by linear-algebra reconstruction recovers these quantities. The determination is among models satisfying these assumptions, including* $r \leq \lfloor C/2 \rfloor$*.*

The state count is recovered from the input; it is not supplied to the construction. The number of tests must nevertheless satisfy the theorem's bound. The statement treats the counterfactual probabilities as exact input and does not assert a finite-sample error bound.

Proof. Starting from λ, the feasible histories can be identified by retaining hao exactly when $\mathrm{P}(o \mid h, a) > 0$. This probability is obtained from the row for h by fixing test action sequences, with the first starting with a, and summing all test observations except that first observation o. Since $\mathbf{Forw}^K$ contains $\mathbf{Forw}^{K-1}$, it has full column rank. Lemma 1 therefore applies and determines the number of states as $r = \operatorname{rank}(\mathsf{CfactM}(C)^{K,L})$. Put $q = \lfloor C/2 \rfloor$, as in the lemma.

Applying the tensor decomposition result. For $0 \le j \le K$, write $\mathcal{H}_j$ for the feasible histories of lengths at most j, including λ, and let $b_h = (\mathbf{Forw}^K_{h,:})^\top$. Divide T by $|\mathcal{H}_K| n^{2Lq}$ to obtain a joint probability tensor for a uniformly chosen history and two groups of tests with independent uniform action sequences. Define

$$w = \frac{1}{|\mathcal{H}_K|} \sum_{h \in \mathcal{H}_K} b_h.$$

Every $[w]_s$ is positive: otherwise nonnegativity would force a zero state-s column in $\mathbf{Forw}^K$, contradicting full column rank. The normalized tensor has latent weights w and state-s factor columns

$$\frac{\mathbf{Forw}^K_{:,s}}{|\mathcal{H}_K|[w]_s}, \qquad n^{-Lq}(\mathsf{cfactBack}(q)^L_{s,:})^\top, \qquad n^{-Lq}(\mathsf{cfactBack}(q)^L_{s,:})^\top.$$

Each column is a probability vector, and all three factor matrices have full column rank by lemma 1. Their pairwise moments are obtained by marginalizing the normalized tensor. The multi-view reduction of Anandkumar et al. (2012, Theorem 3.6), followed by their Theorem 4.3, therefore recovers the third-view columns and weights $[w]_s$, up to a common permutation. Their reduction to the full-rank column spaces covers the possibly different dimensions of the three modes. Multiplying the recovered third-view columns by n^{Lq} gives $(\mathsf{cfactBack}(q)^L)^\top$ in a fixed state order. The normalization is an operation on the supplied probabilities and does not require a new sampling experiment or full support of b_0.

Recovering the backward matrix and history beliefs. In each recovered row of $\mathsf{cfactBack}(q)^L$, fix the action sequences of all tests except the first and sum over their observation sequences. Each omitted test contributes a factor of one, leaving the corresponding row of $\mathbf{Back}^L$. The same marginalization of the supplied matrix, retaining q tests, gives $\mathsf{CfactM}(q)^{K,L}$. Since $\mathsf{cfactBack}(q)^L$ has full row rank, the factorization yields

$$\mathbf{Forw}^K = \mathsf{CfactM}(q)^{K,L}(\mathsf{cfactBack}(q)^L)^\dagger,$$

where † denotes the Moore–Penrose inverse, which is a right inverse of $\mathsf{cfactBack}(q)^L$. This recovers every $b_h = (\mathbf{Forw}^K_{h,:})^\top$, including $b_0 = b_\lambda$, in the state order already selected.

Recovering the observation matrices. Fix an action a, observation o, and any action suffix $g \in \mathcal{A}^{L-1}$. Summing all observations after the first gives

$$O^a_{s,o} = \sum_{y \in \mathcal{O}^{L-1}} \mathbf{Back}^L_{s,(ag,oy)}.$$

Here ag and oy denote concatenation. For $L = 1$, both suffixes are empty. The sum is independent of the chosen action suffix because it marginalizes all observations after the first.

Recovering the transition matrices. For $h \in \mathcal{H}_{K-1}$, let

$$z_{h,a,o} = \mathrm{P}(o \mid h, a) = \sum_{s=1}^{r} [b_h]_s O^a_{s,o}.$$

These probabilities are known from the recovered beliefs and observation matrices. When $z_{h,a,o} > 0$, the history hao belongs to $\mathcal{H}_K$, and its recovered belief is the distribution after the action and observation. For each possible state t reached by that action, the observation-before-transition convention gives

$$z_{h,a,o}[b_{hao}]_t = \sum_{s=1}^{r} [b_h]_s O^a_{s,o} T^a_{s,t}.$$

If $z_{h,a,o} = 0$, the right side is zero by nonnegativity, so there is no need to define a belief for that impossible history. Summing over the observations and using $\sum_o O^a_{s,o} = 1$ therefore gives

$$\sum_{o : z_{h,a,o} > 0} z_{h,a,o} b^\top_{hao} = b^\top_h T^a.$$

For each action a, form the known matrix G^a with the same history-row ordering as $\mathbf{Forw}^{K-1}$ and rows

$$G^a_{h,:} = \sum_{o : z_{h,a,o} > 0} z_{h,a,o} b^\top_{hao}, \qquad h \in \mathcal{H}_{K-1}.$$

The preceding identities give

$$G^a = \mathbf{Forw}^{K-1} T^a. \tag{8}$$

Full column rank makes $(\mathbf{Forw}^{K-1})^\dagger$ a left inverse, so (8) yields

$$T^a = (\mathbf{Forw}^{K-1})^\dagger G^a.$$

This recovers every transition matrix in the same state order as the observation matrices. All recovered matrices equal the true matrices in that order, so no additional projection onto stochastic matrices is needed. □

D Full Algorithm and Omitted Proofs

D.1 The Full Algorithm

D.1.1 Estimating and Filtering the Empirical SDM

First, we recall that $n = |\mathcal{A}|$ is the number of actions, and $m = |\mathcal{O}|$ is the number of observations. The data trajectory is gathered up to N steps, so that $\mathcal{D} = (ao)_{1:N}$. The alphabets are finite and nonempty, and $N \geq 1$. Fix positive integers K, L, C for the maximum history length, test length, and number of counterfactual tests, respectively. The procedure also takes a positive integer frequency-filter threshold $n_{\min}$ and a stopping tolerance $\tau > 0$. For convenience, we will denote some auxilary quantities to abbreviate later expressions in these proofs. A test consists of an L-long string of actions and L-long string of observations. We will often use the letter $g \in \mathcal{A}^L$ to refer to these strings of actions and $y \in \mathcal{O}^L$ for the string of observations. We will use the tuple (g, y) to reference the action and observation strings together. To denote the total number of such strings, we use $G = |\mathcal{A}^L| = n^L$ and $d = |\mathcal{O}^L| = m^L$.

Estimate the SDM row for every history h with $|h| \leq K$ using eq. (6), including the empty history. Retain all complete length-L test windows, including overlaps. For the empty history, the history indicator equals one at every offset; when $|h| + L > N$, the corresponding sums are empty and their counts are zero.

For a given indexing history h, the row of the empirical SDM $\hat{\mathcal{M}}^{K,L}_{h,:}$ can be divided into G blocks, each block corresponding to the frequency of the possible observation sequences to be received under a particular sequence of actions $g \in \mathcal{A}^L$.

SDM Row Norm Structure. Per eq. (6), given a history h and an action sequence $g \in \mathcal{A}^L$, if there is an observation sequence $y \in \mathcal{O}^L$ at $\hat{\mathcal{M}}^{K,L}_{h,(g,y)} > 0$, then $\sum_{y \in \mathcal{O}^L} \hat{\mathcal{M}}^{K,L}_{h,(g,y)} = 1$ (if no observation sequence has been collected for action group g, then we fill that block with zeros). Thus, we can bound the squared norm of all the rows of the empirical SDM as follows since all values of the empirical SDM fall between $[0, 1]$:

$$\begin{aligned}\|\hat{\mathcal{M}}^{K,L}_{h,:}\|_2^2 &= \sum_{g \in \mathcal{A}^L} \sum_{y \in \mathcal{O}^L} \left(\hat{\mathcal{M}}^{K,L}_{h,(g,y)}\right)^2 \\ &\leq \sum_{g \in \mathcal{A}^L} \sum_{y \in \mathcal{O}^L} \hat{\mathcal{M}}^{K,L}_{h,(g,y)} \leq G = \left\|\mathbf{Back}^L_{s,:}\right\|_2^2.\end{aligned}$$

The last equality holds for $\mathbf{Back}^L$ matrices that are deterministic, and for each action sequence $g \in \mathcal{A}^L$ there is only one observation sequence $y \in \mathcal{O}^L$ that has an entry of one, and all other entries in the corresponding block of g are zero.

The History Frequency Filter. We now turn to the row filter, a mechanism that contributes to CfactSP's robustness. Let $N(h, g)$ denote the denominator in eq. (6) for history h and action sequence g:

$$N(h, g) = \sum_{i=0}^{N-(|h|+L)} \mathbf{1}\{(ao)_{i+1:i+|h|} = h\}\mathbf{1}\{a_{i+|h|+1:i+|h|+L} = g\}. \tag{9}$$

This quantity counts the number of times the history h has been observed, followed by the action sequence $g \in \mathcal{A}^L$. We use this count to filter rows under consideration for selection as an anchor row

in successive projection. In particular we choose an integer $n_{\min}$, which may vary with the length of the data trajectory N, to use as a threshold for all g given a history h that indexes a row of the empirical SDM:

$$\min_{g \in \mathcal{A}^L} N(h, g) \geq n_{\min}, \qquad n_{\min} \in \{1, 2, \ldots\}.$$

We vary $n_{\min}$ linearly with the trajectory length N with a frequency rate ν.

$$n_{\min} = \lceil \nu N \rceil, \qquad \nu > 0.$$

Importantly, for a row to be admissible to this filter, every action group block of the empirical SDM must have a positive denominator (no action blocks with all zeros).

For every history h passing the frequency filter, identify the most frequent observation sequence in each action block, which ties broken in a consistent, but arbitrary way:

$$\widehat{y}_h(g) = \arg\max_{y \in \mathcal{O}^L} \hat{\mathcal{M}}^{K,L}_{h,(g,y)}. \tag{10}$$

Let E, the set of histories under consideration for selection in successive projection, contain the histories h with $|h| < K$ that pass the frequency filter and satisfy the following two conditions for every action $a \in \mathcal{A}$. First, the most frequent observation sequences $\widehat{y}_h(g)$ must have the same first observation for every action sequence g beginning with a. Denote this common observation by $o_h(a)$. Second, the history obtained by appending $(a, o_h(a))$ to h must also pass the frequency filter:

$$\min_{g \in \mathcal{A}^L} N\big(h \cdot (a, o_h(a)), g\big) \geq n_{\min} \qquad \text{for every } a \in \mathcal{A}.$$

The restriction $|h| < K$ leaves room for an additional action-observation extension (to be used to reconstruct state transitions). These appended histories need only pass the frequency filter; they need not themselves belong to E. The set E is formed once and kept fixed. Selection chooses among its histories that have not yet been selected. If E is empty, the algorithm returns INCONCLUSIVE.

D.1.2 Successive Projection and Stopping

Applications of the Khatri-Rao Product and Successive Projection. For the chosen number C of counterfactual tests, we take the C-fold row-wise Khatri–Rao product of a row vector $z \in \mathbb{R}^{Gd}$ after dividing it by $\sqrt{G}$, denoted by $\Phi_C(\cdot)$:

$$\Phi_C(z) = \left(\frac{z}{\sqrt{G}}\right)^{\odot C}. \tag{11}$$

Here $\odot$ denotes the row-wise Khatri–Rao product, with one factor for each counterfactual test. At an exact uncorrupted anchor, this gives the corresponding counterfactual backward row from eq. (5), scaled by $G^{-C/2}$. For a general mixture row, the product need not give its true counterfactual distribution. In our analysis, every row is normalized by $1/\sqrt{G}$. Each uncorrupted backward row therefore has unit norm before taking the product, while differences in empirical row norms are retained for the perturbation bounds used in selection.

Initialization. Initialize an empty list F of selected rounded rows and an empty list of corresponding selected histories. Initially, the set of histories still available for selection is $\mathcal{H}_{\text{active}} = E$, and R is the identity matrix on $\mathbb{R}^{(Gd)^C}$. As rows are selected, R projects onto the orthogonal complement of the span of their Khatri–Rao products, expressed as columns.

Stopping Condition. The stopping tolerance is an input $\tau > 0$. The recovery guarantee uses

$$\tau = \frac{1}{2\sqrt{G}}.$$

Before each selection, form the convex hull of the origin and the rounded rows selected so far:

$$\mathcal{C}_F = \operatorname{conv}(\{0\} \cup \{f : f \in F\}).$$

The stopping condition is

$$\max_{h \in E} \operatorname{dist}_2(\hat{\mathcal{M}}^{K,L}_{h,:}, \mathcal{C}_F) \leq \tau. \tag{12}$$

This test uses every history in the fixed set E, including those already selected. If the condition holds and F is nonempty, proceed to the matrix construction in section D.1.3. If it holds while F is empty, return INCONCLUSIVE. If the condition fails and no history remains in $\mathcal{H}_{\text{active}}$, also return INCONCLUSIVE.

Selecting the Next History. When the stopping condition fails and candidates remain, choose

$$h^\star = \operatorname*{arg\,max}_{h \in \mathcal{H}_{\text{active}}} \|R\,\Phi_C(\hat{\mathcal{M}}^{K,L}_{h,:})^\top\|_2,$$

using a fixed ordering of histories to resolve ties. The projected norm is computed from the empirical row. The following kernel formula evaluates it using inner products of the original rows.

Kernel Evaluation. The C-fold Khatri–Rao products have $(Gd)^C$ entries, but their inner products can be computed from the original SDM rows. Each of the C factors contributes the same inner product, giving

$$\langle \Phi_C(z), \Phi_C(z')\rangle = \left(\frac{\langle z, z'\rangle}{G}\right)^C. \tag{13}$$

This identity provides a kernel for evaluating the projected norms used in successive projection without constructing the Khatri–Rao products explicitly.

Let F be the current list of selected rounded rows. Define their Gram matrix Q_F and, for a candidate row z, the column vector $v_F(z)$ by the following entries, indexed by $f, f' \in F$:

$$(Q_F)_{f,f'} = \left(\frac{\langle f, f'\rangle}{G}\right)^C,$$
$$[v_F(z)]_f = \left(\frac{\langle f, z\rangle}{G}\right)^C.$$

The matrix R projects onto the orthogonal complement of the span of the columns $\Phi_C(f)^\top$, $f \in F$. Subtracting the squared norm of the projection onto this span from the squared norm of $\Phi_C(z)$ gives

$$\|R\,\Phi_C(z)^\top\|_2^2 = \left(\frac{\|z\|_2^2}{G}\right)^C - v_F(z)^\top Q_F^\dagger v_F(z), \tag{14}$$

where $Q_F^\dagger$ is the Moore–Penrose pseudoinverse. When F is empty, the subtracted term is zero. Substituting $z = \hat{\mathcal{M}}^{K,L}_{h,:}$ gives the squared projected norm used to choose the next history. Thus the projected norm compares an empirical row with the span determined by the selected rounded rows, using only inner products of vectors with Gd entries.

Rounding and Updating the Projection. After selecting $h^\star$, round its empirical SDM row. For any history h passing the frequency filter, the rounding rule is

$$\mathcal{R}(h)_{(g,y)} = \mathbf{1}\{y = \widehat{y}_h(g)\}. \tag{15}$$

The rounded row has exactly one entry equal to one in each action block, at the most frequent observation sequence defined in eq. (10). Apply this rule to the selected history and project the Khatri–Rao product of the resulting row:

$$f^\star = \mathcal{R}(h^\star), \qquad \mathbf{q} = R\,\Phi_C(f^\star)^\top.$$

If $f^\star$ already occurs in F or $\|\mathbf{q}\|_2 = 0$, return INCONCLUSIVE. Otherwise, update the projection matrix by

$$R \leftarrow R - \frac{\mathbf{q}\mathbf{q}^\top}{\|\mathbf{q}\|_2^2}.$$

The corresponding kernel update enlarges the Gram matrix to include $f^\star$. Write $F_+ = (F, f^\star)$ for the enlarged list. Since the rounded row has one entry equal to one in each of its G blocks, $\|f^\star\|_2^2 = G$ and $\|\Phi_C(f^\star)\|_2^2 = 1$. Therefore,

$$Q_{F_+} = \begin{pmatrix} Q_F & v_F(f^\star) \\ v_F(f^\star)^\top & 1 \end{pmatrix}, \qquad v_{F_+}(z) = \begin{pmatrix} v_F(z) \\ (\langle f^\star, z\rangle/G)^C \end{pmatrix}.$$

The existing block Q_F stays unchanged because its entries are inner products of the selected Khatri–Rao product columns before projection. Recompute the pseudoinverse of the enlarged matrix and use Q_{F_+} and $v_{F_+}(z)$ in eq. (14) to compute subsequent squared projected norms. This accounts for the new projection through the enlarged selected span.

Append $h^\star$ and $f^\star$ to the selected-history and rounded-row lists, and remove $h^\star$ from $\mathcal{H}_{\text{active}}$. The vector $\mathbf{q}$ is orthogonal to the previously selected product columns, so the update removes its direction and preserves the stated projection property. Return to the stopping test in eq. (12) before selecting another history. Thus the projected norms used for selection are computed from empirical rows, projection updates use rounded rows, and stopping is tested on the original SDM rows.

D.1.3 Estimating the States and Matrices

When the stopping condition is met with F nonempty, the estimated number of states is $\widehat{r} = |F|$. Write the selected histories as $(h_1, \ldots, h_{\widehat{r}})$ and the corresponding rounded rows as $F = (f_1, \ldots, f_{\widehat{r}})$. For each estimated state i and action $a \in \mathcal{A}$, form the indicated extension $h_i^a = h_i \cdot (a, o_{h_i}(a))$, round its empirical row using eq. (15), and find its matching rounded rows:

$$J_{i,a} = \{j \in \{1, \ldots, \widehat{r}\} : \mathcal{R}(h_i^a) = f_j\}.$$

The construction of E ensures that each h_i^a passes the frequency filter, so every block has observations on which to base the rounding. If any $J_{i,a}$ has anything other than one element, return INCONCLUSIVE. Otherwise, let $j(i,a)$ denote its unique element and set

$$\widehat{T}_{ij}^a = \mathbf{1}\{j = j(i,a)\}, \qquad \widehat{O}_{ib}^a = \mathbf{1}\{o^b = o_{h_i}(a)\},$$

for every $i, j \in \{1, \ldots, \widehat{r}\}$, $a \in \mathcal{A}$, and $b \in \{1, \ldots, m\}$, where $\mathcal{O} = \{o^1, \ldots, o^m\}$.

The procedure returns the estimated state set $\widehat{\mathcal{S}} = \{\widehat{s}^1, \ldots, \widehat{s}^{\widehat{r}}\}$ and deterministic matrices

$$\widehat{T}^a \in \{0,1\}^{\widehat{r} \times \widehat{r}}, \qquad \widehat{O}^a \in \{0,1\}^{\widehat{r} \times m}, \qquad a \in \mathcal{A}.$$

Each matrix row has one entry equal to one. The selected histories and their rounded backward rows are retained with the output. The cases specified above return INCONCLUSIVE instead of a model.

D.2 Recovery Guarantee and Notation

D.2.1 Assumptions and Repeating Quantities

Let $\mathcal{S} = \{s^1, \ldots, s^r\}$ be the state set of the uncorrupted POMDP, whose transition and observation matrices T^a and O^a are deterministic. The data $\mathcal{D} = (ao)_{1:N}$ come from a fixed ϵ-corrupted POMDP $\widetilde{\mathcal{P}}$ as defined in definition 3, with an arbitrary initial belief. Actions are sampled independently and uniformly from $\mathcal{A}$, independently of the initial state and the environment's random draws.

Corruption Bound. The transition and observation row errors are each at most ϵ in total variation. For a state s^i and action a, let s^j and o^b be their unique uncorrupted successor and observation, so that $T_{ij}^a = O_{ib}^a = 1$. The total variation distance from a point mass is the probability assigned to other outcomes. A union bound therefore gives

$$\begin{aligned} \mathbb{P}\big\{(o_t, s_{t+1}) \neq (o^b, s^j) \mid s_t = s^i, a_t = a\big\} &\leq (1 - \widetilde{O}_{ib}^a) + (1 - \widetilde{T}_{ij}^a) \\ &\leq 2\epsilon = \beta. \end{aligned} \tag{16}$$

We assume $0 \leq \epsilon < 1/2$, so $\beta < 1$. The proofs use this uniform bound on the probability that either the observation or the successor differs from the uncorrupted outcome. Any smaller uniform bound on that joint probability may be used in place of β. Deviations at different visits need not be independent. We continue to use λ for the empty history and reserve $\delta \in (0,1)$ for the failure probability.

Connectivity and State Separation. Assume the uncorrupted transition graph is strongly connected. Let D be its directed diameter: the maximum, over ordered pairs of states, of the shortest

action-sequence length taking the first state to the second. The test length L must distinguish every pair of states: for any $s^i \neq s^j$, some action sequence $g \in \mathcal{A}^L$ produces different observation sequences when started from those two states in the uncorrupted model. Equivalently,

$$\mathbf{Back}^L_{s^i,:} \neq \mathbf{Back}^L_{s^j,:} \qquad \text{for all } i \neq j.$$

History Length. An anchor sequence for a state s is an action-observation sequence that can occur in the uncorrupted model and for which every uncorrupted execution producing that sequence ends at s, regardless of its starting state. Let K_* be the smallest length bound such that every state has an anchor sequence of length at most K_*, allowing the empty sequence. We assume

$$K \geq K_* + 1. \tag{17}$$

Thus each state has an anchor sequence satisfying the selection rule's length restriction $|h| < K$, and its one-step extensions have length at most K. The extra step is needed for the histories used to reconstruct state transitions. Under the connectivity and state-separation assumptions, lemma 4 gives $K_* \leq (r-1)L+D$, so $K \geq (r-1)L+D+1$ is sufficient for this history-length requirement.

D.2.2 POMDP-Dependent Quantities

The recovery bounds depend on how well the uncorrupted states can be distinguished, how often identifying histories occur, and the accuracy of the empirical SDM rows. We introduce the quantities for these bounds here, taking $N \geq \max\{2, L+1\}$ and $\delta \in (0,1)$.

Separation of Backward Rows. Recall that $G = n^L$ and $d = m^L$ count the length-L action and observation sequences, respectively. Each uncorrupted backward row has exactly one entry equal to one in each of its G action blocks, so $\|\mathbf{Back}^L_{s,:}\|_2^2 = G$. For distinct states $s, t \in \mathcal{S}$, define

$$c_{st} = \frac{\langle \mathbf{Back}^L_{s,:}, \mathbf{Back}^L_{t,:} \rangle}{G}, \qquad \widehat{c} = \max_{s \neq t} c_{st}, \qquad u = 1 - \widehat{c}. \tag{18}$$

The overlap c_{st} is the fraction of action sequences that produce the same observation sequence from s and t in the uncorrupted model. State separation ensures that at least one action sequence distinguishes each pair, giving

$$0 \leq \widehat{c} \leq 1 - \frac{1}{G}, \qquad \frac{1}{G} \leq u \leq 1.$$

When $r = 1$, set $\widehat{c} = 0$ and $u = 1$. Thus u bounds below the fraction of action sequences distinguishing any pair of states. By eq. (13), the corresponding inner product after taking the normalized C-fold Khatri–Rao products is c_{st}^C. These overlaps provide quantitative bounds on the separation of the Khatri–Rao product rows and the norms of products of mixture rows. The recovery guarantee uses these bounds to give a sufficient number of counterfactual tests for robust successive projection.

Fixed Anchor Histories and Their Extensions. For each state s, fix an anchor history $h_s^\star$ of length at most K_*, independently of the sampled trajectory. Let $\mathcal{H}$ be the set consisting of these histories and all extensions $h_s^\star \cdot (a, o^b)$ with $a \in \mathcal{A}$ and $O^a_{s,b} = 1$. Here the matrix row indexed by s is the row corresponding to that state, and b ranges over the observation indices. Each extension identifies the uncorrupted successor of s under action a. Since $K \geq K_* + 1$, this fixed family satisfies

$$|\mathcal{H}| \leq r(n+1), \qquad |h| \leq K \quad \text{for every } h \in \mathcal{H}. \tag{19}$$

The later occurrence bounds will ensure that every history in this family passes the frequency filter with high probability.

Simultaneous Concentration Bounds. To cover every empirical row that the algorithm might use, the probability bounds range over all action-observation histories of length at most K. Define

$$\begin{aligned} J_N &= 1 + \lceil \log_2 N \rceil, \\ F_0 &= GL(d+1) \sum_{j=0}^{K} (nm)^j + 2rn \sum_{j=0}^{K-1} (nm)^j, \\ \Lambda &= \log \frac{4 J_N F_0}{\delta}. \end{aligned} \tag{20}$$

All logarithms except $\log_2$ are natural. The first term in F_0 counts one action-block count and d observation sequence counts for every history and action sequence, with a factor of L for separating test starting positions by their remainder modulo L. This separation is used in section D.3.4 to handle overlapping test windows. The second term reserves two additional martingale sums for every history of length less than K, state, and action, as needed for the parent–child argument in section D.7.2. The factor J_N accounts for the variance ranges in lemma 2. With the stated choice of Λ, a failure bound of $2J_N e^{-\Lambda}$ for each of the F_0 sums gives a total bound of $\delta/2$.

Empirical Row Error. Define the test-corruption term and the two empirical-row error quantities by

$$\begin{aligned} \eta_\beta &= 1 - (1-\beta)^L, \\ \eta(n_{\min}) &= 2(\sqrt{d}+1)\sqrt{\frac{L\Lambda}{n_{\min}}} + \eta_\beta, \\ \zeta(n_{\min}) &= 8\sqrt{\frac{L\Lambda}{n_{\min}}} + \eta_\beta. \end{aligned} \tag{21}$$

The term η_β bounds the probability of at least one joint deviation during a length-L test. For rows passing the frequency filter with $n_{\min} \geq 64L\Lambda$, theorem 5 will bound the full-row Euclidean error by $2\sqrt{G}\,\eta(n_{\min})$ and each coordinate error by $\zeta(n_{\min})$ on a simultaneous probability event. The reference rows are mixtures of uncorrupted backward rows weighted by the empirical frequencies of the states reached after the counted action-observation sequences equal to each history, as defined in section D.3.1. The square-root terms account for sampling error, while η_β accounts for corrupted test responses.

Occurrences of the Fixed Histories. To produce a given history in $\mathcal{H}$, an uncorrupted execution can first reach a state capable of generating it in at most D steps, then follow the history in at most K steps. Set

$$W = D + K, \qquad \alpha_\beta = n^{-W}(1-\beta)^W, \qquad \mu = \left\lfloor \frac{N-L}{W} \right\rfloor \alpha_\beta.$$

Here $W \geq 1$, and α_β is a uniform lower bound on the conditional probability of an uncorrupted occurrence of a fixed history during a block of W steps. The floor in the definition of μ counts disjoint blocks with at least L trajectory steps remaining for a complete subsequent test. Multiplying this count by α_β gives the deterministic quantity μ used in the occurrence bound. On the event established in lemma 12, every history in $\mathcal{H}$ has at least $\mu/2$ uncorrupted occurrences with complete test windows.

Perturbation of the Supplied Histories. Define

$$\Lambda_c = \log\frac{4}{\delta}, \qquad B_N = \begin{cases} 0, & \beta = 0, \\ 2\beta N + \Lambda_c, & \beta > 0. \end{cases}$$

Lemma 13 bounds the total number of joint deviations in the trajectory by B_N with probability at least $1 - \delta/4$. When $\mu > 0$, define

$$\kappa_N = \frac{2KB_N}{\mu}.$$

For any history in $\mathcal{H}$, each joint deviation can affect at most K occurrences of that history. Combining the bound B_N with at least $\mu/2$ uncorrupted occurrences bounds the fraction of its occurrences counted by the estimator that end outside the identified state by κ_N. This bounds the perturbation of the empirical distributions of states reached after the counted action-observation sequences equal to the supplied histories in $\mathcal{H}$. The quantities $\eta(n_{\min})$ and $\zeta(n_{\min})$ separately bound the errors in estimating test responses. At $\beta = 0$, both η_β and κ_N are zero whenever $\mu > 0$.

D.3 Concentration Along a Trajectory Without Reset

D.3.1 Finite-Record Reference Rows

We compare each empirical SDM row with a mixture of uncorrupted backward rows whose weights describe the latent states following the observed occurrences of its indexing history. The same weights will be used for every action block of that row.

History Occurrences. Fix a history h of length $t = |h| \le K$ and use the window offset i from eq. (6). The history occupies positions $i+1 : i+t$, and its test occupies positions $i+t+1 : i+t+L$. Define the indicator that the recorded action-observation sequence at these steps equals h, and the number of counted occurrences, by

$$H_i(h) = \mathbf{1}\{(ao)_{i+1:i+t} = h\}, \qquad N_h = \sum_{i=0}^{N-(t+L)} H_i(h).$$

Only occurrences with a complete subsequent test window contribute to N_h, and overlapping occurrences are retained. A sum with a negative upper limit is zero. Equivalently, write the set of steps at which the counted sequences equal to h finish as

$$\mathcal{I}_h = \{j : t \le j \le N - L,\ (ao)_{j-t+1:j} = h\}, \qquad N_h = |\mathcal{I}_h|.$$

A nonempty sequence beginning at step $i+1$ finishes at step $j = i+t$, and the state reached after that sequence is $s_{j+1} = s_{i+t+1}$. For the empty history, each index $0 \le j \le N - L$ contributes one counted occurrence. Here $j = 0$ denotes the time before the first action, when the state is s_1, and $j > 0$ denotes the time immediately after step j. If the index range is empty, then $\mathcal{I}_h = \varnothing$.

Each counted occurrence is followed by exactly one length-L action sequence, so the action-block counts in eq. (9) satisfy

$$N_h = \sum_{g \in \mathcal{A}^L} N(h, g). \tag{22}$$

In particular, every history passing the frequency filter has $N_h \ge G n_{\min} > 0$.

Finite-Record Forward Weights. Let $\overline{\mathbf{Forw}}^K$ be the forward matrix that counts the finite-sample empirical distribution of state visitations for each indexing history. For a history with $N_h > 0$, define

$$\begin{aligned} \overline{\mathbf{Forw}}^K_{h,s} &= \frac{1}{N_h} \sum_{i=0}^{N-(t+L)} H_i(h) \mathbf{1}\{s_{i+t+1} = s\} \\ &= \frac{1}{|\mathcal{I}_h|} \sum_{j \in \mathcal{I}_h} \mathbf{1}\{s_{j+1} = s\}. \end{aligned} \tag{23}$$

These are state frequencies in the finite record, distinct from the population forward matrix $\mathbf{Forw}^K$. They are used in the proof and are not computed by the algorithm. The definition averages over all counted occurrences of h, regardless of the subsequent action sequence. The weights are nonnegative and sum to one over $s \in \mathcal{S}$. They are defined only when $N_h > 0$; the zero-fill convention for empirical SDM blocks remains as specified in section D.1.1.

The Common Reference Mixture. For each history with $N_h > 0$, the reference row is

$$\begin{aligned} \overline{\mathbf{Forw}}^K_{h,:} \mathbf{Back}^L &= \sum_{s \in \mathcal{S}} \overline{\mathbf{Forw}}^K_{h,s} \mathbf{Back}^L_{s,:} \\ &= \frac{1}{N_h} \sum_{i=0}^{N-(t+L)} H_i(h) \mathbf{Back}^L_{s_{i+t+1},:}. \end{aligned}$$

Thus each block of the reference row is a probability vector, and

$$\left\| \overline{\mathbf{Forw}}^K_{h,:} \mathbf{Back}^L \right\|_2 \le \sum_{s \in \mathcal{S}} \overline{\mathbf{Forw}}^K_{h,s} \|\mathbf{Back}^L_{s,:}\|_2 = \sqrt{G}.$$

An empirical action block uses only the occurrences followed by its particular action sequence, while its reference block uses the common weights in eq. (23). The concentration argument in section D.3.4 bounds the resulting sampling differences together with the deviations of the random action-block counts. It also accounts for the difference between corrupted test responses and the uncorrupted backward rows. The reference mixture is defined from the actual latent states in the finite trajectory, so this comparison applies from an arbitrary initial belief without a stationarity assumption. The simultaneous bounds use the concentration quantities already defined in eq. (20).

D.3.2 SCALAR CONCENTRATION

The centered count increments used below are dependent, but have conditional mean zero given the information available before the next test. We use Freedman's inequality to bound their sum in terms of their accumulated conditional variance. The variance bound in our application depends on the random number of history occurrences. To accommodate this randomness, we apply the scalar bound simultaneously at a finite set of deterministic variance limits.

Lemma 2 (Bounded increments and variance peeling)**.** *Let $(\mathcal{G}_i)_{i=0}^{q}$ be a filtration with a deterministic terminal index q. For $i = 0, \dots, q-1$, suppose that X_i is $\mathcal{G}_{i+1}$-measurable, $\mathbb{E}[X_i \mid \mathcal{G}_i] = 0$, and $|X_i| \le 1$ almost surely. Define the terminal sum and its predictable variance by*

$$M = \sum_{i=0}^{q-1} X_i, \qquad V = \sum_{i=0}^{q-1} \mathbb{E}[X_i^2 \mid \mathcal{G}_i].$$

For the concentration parameter $\Lambda > 1$ in eq. (20) *and every deterministic $v \ge 0$,*

$$\mathbb{P}\left\{ |M| \ge \sqrt{2v\Lambda} + \frac{2}{3}\Lambda,\ V \le v \right\} \le 2e^{-\Lambda}. \tag{24}$$

If w is a possibly random variable satisfying $V \le w \le N$ almost surely, then

$$\mathbb{P}\left\{ |M| > 2\sqrt{w\Lambda} + 3\Lambda \right\} \le 2J_N e^{-\Lambda}, \qquad J_N = 1 + \lceil \log_2 N \rceil. \tag{25}$$

Proof. For $a, v > 0$, Freedman's inequality, in the form stated by Tropp (2011, Theorem 1.1), applied to the partial sums of X_i and $-X_i$, gives

$$\mathbb{P}\{|M| \ge a, V \le v\} \le 2 \exp\left\{ -\frac{a^2}{2(v + a/3)} \right\}.$$

This bound also holds at $v = 0$ by taking $v \downarrow 0$. Substituting $a = \sqrt{2v\Lambda} + 2\Lambda/3$ proves eq. (24).

For the random variance bound, apply eq. (24) at each of the J_N deterministic limits

$$v_j = 2^j, \qquad j = 0, \dots, \lceil \log_2 N \rceil.$$

A union bound gives an event of probability at least $1 - 2J_N e^{-\Lambda}$ on which every implication

$$V \le v_j \quad \Longrightarrow \quad |M| < \sqrt{2v_j\Lambda} + \frac{2}{3}\Lambda$$

holds simultaneously. On this event, if $w \ge 1$, choose the smallest v_j with $v_j \ge w$. Then $V \le w \le v_j \le 2w$, and hence

$$|M| < 2\sqrt{w\Lambda} + \frac{2}{3}\Lambda \le 2\sqrt{w\Lambda} + 3\Lambda.$$

If $0 \le w < 1$, use $v_0 = 1$ and $\Lambda > 1$ to obtain

$$|M| < \sqrt{2\Lambda} + \frac{2}{3}\Lambda \le 3\Lambda \le 2\sqrt{w\Lambda} + 3\Lambda.$$

This includes $w = 0$ and proves eq. (25). The choice of the covering limit is made on the simultaneous event, so the argument does not condition on the realized value of w. □

D.3.3 CORRUPTED TEST RESPONSES

Let $\widetilde{\mathbf{Back}}^{L}$ be the backward matrix of the fixed corrupted POMDP. Its entry $\widetilde{\mathbf{Back}}^{L}_{s,(g,y)}$ is the probability of observing $y \in \mathcal{O}^L$ when executing the action sequence $g \in \mathcal{A}^L$ from state s. For a row vector z, write $z|_g$ for its block of coordinates indexed by (g, y) as y ranges over $\mathcal{O}^L$. Each block of $\widetilde{\mathbf{Back}}^{L}$ is a probability vector, while each corresponding block of $\mathbf{Back}^{L}$ is a point mass. We use $\|p - q\|_{\mathrm{TV}} = \frac{1}{2}\|p - q\|_1$ for the total variation distance between probability vectors.

Lemma 3 (Test-window displacement)**.** *For every state $s \in \mathcal{S}$ and action sequence $g \in \mathcal{A}^L$,*

$$\left\| \widetilde{\mathbf{Back}}^{L}_{s,:}|_g - \mathbf{Back}^{L}_{s,:}|_g \right\|_{\mathrm{TV}} \le \eta_\beta = 1 - (1-\beta)^L \le L\beta.$$

Consequently, for every history h with $N_h > 0$ and every $g \in \mathcal{A}^L$, the finite-record weights in eq. (23) *satisfy*

$$\begin{aligned} \left\| \left[\overline{\mathbf{Forw}}^{K}_{h,:} \left(\widetilde{\mathbf{Back}}^{L} - \mathbf{Back}^{L} \right) \right] |_g \right\|_{\infty} &\le \eta_\beta, \\ \left\| \left[\overline{\mathbf{Forw}}^{K}_{h,:} \left(\widetilde{\mathbf{Back}}^{L} - \mathbf{Back}^{L} \right) \right] |_g \right\|_{2} &\le 2\eta_\beta. \end{aligned} \tag{26}$$

Proof. Fix s and g, and compare the corrupted and uncorrupted executions from s under this same action sequence. Let E_ℓ be the event that their observations and successor states agree at every one of the first ℓ steps, with E_0 the sure event. As long as the executions agree, their current states coincide, so eq. (16) bounds the conditional probability of a deviation at the next step by β. Sequential conditioning therefore gives

$$\mathbb{P}(E_\ell) \ge (1-\beta)\mathbb{P}(E_{\ell-1}), \qquad 1 \le \ell \le L, \qquad \mathbb{P}(E_L) \ge (1-\beta)^L.$$

This uses conditional bounds and does not require independent deviations. Let $y^\star$ be the unique observation sequence produced by the uncorrupted execution. On E_L, the corrupted execution also produces $y^\star$, so its distance from the uncorrupted point mass is

$$\left\| \widetilde{\mathbf{Back}}^{L}_{s,:}|_g - \mathbf{Back}^{L}_{s,:}|_g \right\|_{\mathrm{TV}} = 1 - \widetilde{\mathbf{Back}}^{L}_{s,(g,y^\star)} \le 1 - \mathbb{P}(E_L) \le \eta_\beta.$$

The inequality $1-(1-\beta)^L \le L\beta$ follows from Bernoulli's inequality.

For $N_h > 0$, the weights $\overline{\mathbf{Forw}}^{K}_{h,s}$ are nonnegative and sum to one, so convexity gives

$$\left\| \left[\overline{\mathbf{Forw}}^{K}_{h,:} \left(\widetilde{\mathbf{Back}}^{L} - \mathbf{Back}^{L} \right) \right] |_g \right\|_{\mathrm{TV}} \le \sum_{s \in \mathcal{S}} \overline{\mathbf{Forw}}^{K}_{h,s} \left\| \widetilde{\mathbf{Back}}^{L}_{s,:}|_g - \mathbf{Back}^{L}_{s,:}|_g \right\|_{\mathrm{TV}} \le \eta_\beta.$$

For a difference of probability vectors, each coordinate has magnitude at most the total variation distance, and the Euclidean norm is at most the ℓ_1 norm, which equals twice that distance. These observations give eq. (26) for every realization of the finite-record weights. □

D.3.4 Direct Counts and the Common Reference Mixture

Fix a history h of length $t \le K$ and an action sequence $g \in \mathcal{A}^L$. We bound deviations of both the observation sequence counts in this action block and its denominator $N(h,g)$. This will compare the empirical block with the corrupted reference mixture $\left(\overline{\mathbf{Forw}}^{K}_{h,:} \widetilde{\mathbf{Back}}^{L} \right) |_g$, using the same finite-record weights for every action sequence.

Centering the Counts. Let

$$\mathcal{F}_j = \sigma(s_1, a_1, o_1, \ldots, s_j, a_j, o_j, s_{j+1}), \qquad 1 \le j \le N, \qquad \mathcal{F}_0 = \sigma(s_1). \tag{27}$$

This filtration includes the latent states and describes the information through action-observation pair j, before action a_{j+1}. For an admissible window offset $0 \le i \le N-(t+L)$, both $H_i(h)$ and the state s_{i+t+1} at the start of the test are $\mathcal{F}_{i+t}$-measurable. The independent uniform actions and the fixed corrupted transition and observation laws give

$$\mathbb{P}\left\{ a_{i+t+1:i+t+L} = g,\ o_{i+t+1:i+t+L} = y \mid \mathcal{F}_{i+t} \right\} = \frac{1}{G} \widetilde{\mathbf{Back}}^{L}_{s_{i+t+1},(g,y)}$$

for every $y \in \mathcal{O}^L$. Similarly, the conditional probability of the action sequence g alone is $1/G$. Define the centered increments for each observation sequence count and for the action-block count by

$$\begin{aligned} X_{i,y} &= H_i(h) \left[\mathbf{1}\{ a_{i+t+1:i+t+L} = g,\ o_{i+t+1:i+t+L} = y \} - \frac{1}{G} \widetilde{\mathbf{Back}}^{L}_{s_{i+t+1},(g,y)} \right], \\ X_{i,\mathrm{den}} &= H_i(h) \left[\mathbf{1}\{ a_{i+t+1:i+t+L} = g \} - \frac{1}{G} \right]. \end{aligned}$$

Each increment has conditional mean zero given $\mathcal{F}_{i+t}$ and absolute value at most one. Since the variance of a Bernoulli variable is at most its success probability, their conditional variances satisfy

$$\mathbb{E}[X_{i,y}^2 \mid \mathcal{F}_{i+t}] \le \frac{H_i(h)}{G}, \qquad \mathbb{E}[X_{i,\mathrm{den}}^2 \mid \mathcal{F}_{i+t}] \le \frac{H_i(h)}{G}. \tag{28}$$

Write the total centered counts as

$$M_y = \sum_{i=0}^{N-(t+L)} X_{i,y}, \qquad M_{\mathrm{den}} = \sum_{i=0}^{N-(t+L)} X_{i,\mathrm{den}} = N(h,g) - \frac{N_h}{G}. \tag{29}$$

These sums include zero increments at offsets where h does not occur, and an empty sum is zero.

Overlapping Test Windows. An increment at offset i is revealed by $\mathcal{F}_{i+t+L}$, after all L test steps have occurred. Partition the admissible offsets according to their remainder $j \in \{0, \dots, L-1\}$ modulo L. For each nonempty class, list its offsets as $i_q = j + qL$, $q = 0, \dots, q_j - 1$, where q_j is determined by N, t, L, and j. Use the filtration

$$\mathcal{G}_q^{(j)} = \mathcal{F}_{j+qL+t}, \qquad q = 0, \dots, q_j.$$

The final index $j + q_j L + t$ is the revealing time for the last test in the class and is at most N. Each increment at i_q is therefore centered given $\mathcal{G}_q^{(j)}$ and measurable with respect to $\mathcal{G}_{q+1}^{(j)}$. Its partial sums form a martingale with a deterministic number of increments, as required by lemma 2.

Let $M_y^{(j)}$ and $M_{\mathrm{den}}^{(j)}$ denote the sums in eq. (29) restricted to class j, and let

$$N_h^{(j)} = \sum_{\substack{0 \le i \le N-(t+L) \\ i \equiv j \pmod L}} H_i(h).$$

By eq. (28), the predictable variance of each class sum is at most $N_h^{(j)}/G \le N$. Applying lemma 2 to all $L(d+1)$ class sums gives, except on an event of probability at most $2L(d+1)J_N e^{-\Lambda}$ for this fixed pair (h, g),

$$|M_y^{(j)}|,\ |M_{\mathrm{den}}^{(j)}| \le 2\sqrt{\frac{N_h^{(j)}\Lambda}{G}} + 3\Lambda \qquad \text{for all } j \text{ and } y.$$

Since $\sum_j N_h^{(j)} = N_h$, Cauchy–Schwarz gives $\sum_j \sqrt{N_h^{(j)}} \le \sqrt{LN_h}$. Summing the class bounds therefore yields

$$|M_y|,\ |M_{\mathrm{den}}| \le b_h, \qquad b_h = 2\sqrt{\frac{LN_h\Lambda}{G}} + 3L\Lambda \qquad \text{for all } y \in \mathcal{O}^L. \tag{30}$$

The random counts enter through the variance bound in lemma 2; we have not conditioned on their realized values.

Bounding the Denominator. Work on the event in eq. (30). If $N_h/G \ge 64L\Lambda$, then

$$b_h \le \left(\frac{1}{4} + \frac{3}{64}\right)\frac{N_h}{G} < \frac{N_h}{2G},$$

and the denominator identity in eq. (29) gives

$$N(h,g) \ge \frac{N_h}{G} - b_h \ge \frac{N_h}{2G} > 0.$$

In particular, if h passes the history frequency filter with $n_{\min} \ge 64L\Lambda$, then eq. (22) implies $N_h/G \ge n_{\min}$, so this lower bound applies. Using $N_h/G \le 2N(h,g)$ and $N(h,g) \ge n_{\min}$, we obtain

$$\frac{b_h}{N(h,g)} \le 2\sqrt{\frac{2L\Lambda}{N(h,g)}} + \frac{3L\Lambda}{N(h,g)} \le 4\sqrt{\frac{L\Lambda}{n_{\min}}}. \tag{31}$$

The last inequality uses $\sqrt{L\Lambda/N(h,g)} \le 1/8$ and $2\sqrt{2} + 3/8 < 4$.

The Count Reference Mixture and the Count Ratio.

For $N(h,g) > 0$, grouping the centering terms by their latent state in eq. (23) gives

$$M_y = N(h,g)\hat{\mathcal{M}}^{K,L}_{h,(g,y)} - \frac{N_h}{G}\left(\overline{\mathbf{Forw}}^K_{h,:}\widetilde{\mathbf{Back}}^L\right)_{(g,y)}.$$

Let $\boldsymbol{M} = (M_y)_{y\in\mathcal{O}^L}$ be the row vector of centered observation sequence counts. Substituting $N_h/G = N(h,g) - M_{\mathrm{den}}$ and rearranging yields the exact identity

$$\hat{\mathcal{M}}^{K,L}_{h,:}|_g - \left(\overline{\mathbf{Forw}}^K_{h,:}\widetilde{\mathbf{Back}}^L\right)|_g = \frac{\boldsymbol{M} - M_{\mathrm{den}}\left(\overline{\mathbf{Forw}}^K_{h,:}\widetilde{\mathbf{Back}}^L\right)|_g}{N(h,g)}. \tag{32}$$

The reference block is a probability vector, so its coordinate and Euclidean norms are at most one. The count bounds give $\|\boldsymbol{M}\|_\infty \le b_h$ and $\|\boldsymbol{M}\|_2 \le \sqrt{d}\,b_h$. Thus, for a history passing the frequency filter with $n_{\min} \ge 64L\Lambda$, eqs. (31) and (32) imply

$$\begin{aligned}\left\|\hat{\mathcal{M}}^{K,L}_{h,:}|_g - \left(\overline{\mathbf{Forw}}^K_{h,:}\widetilde{\mathbf{Back}}^L\right)|_g\right\|_\infty &\le \frac{2b_h}{N(h,g)} \le 8\sqrt{\frac{L\Lambda}{n_{\min}}},\\ \left\|\hat{\mathcal{M}}^{K,L}_{h,:}|_g - \left(\overline{\mathbf{Forw}}^K_{h,:}\widetilde{\mathbf{Back}}^L\right)|_g\right\|_2 &\le \frac{(\sqrt{d}+1)b_h}{N(h,g)} \le 4(\sqrt{d}+1)\sqrt{\frac{L\Lambda}{n_{\min}}}.\end{aligned} \tag{33}$$

The weights $\overline{\mathbf{Forw}}^K_{h,:}$ average over all counted occurrences of h and are shared by every action block. The next subsection makes these sampling bounds simultaneous over all candidate histories and action sequences, then combines them with lemma 3 to compare with the uncorrupted reference rows.

D.3.5 Simultaneous Row Bounds

We now combine the count bounds over all histories and action sequences with the test-corruption bound. The probability event must also include the one-step counts used later to recover transitions from the histories selected by the algorithm.

For a fixed history h of length $t < K$, state $s \in \mathcal{S}$, and action $a \in \mathcal{A}$, let $(o^\star, s^\star)$ be the uncorrupted observation and successor from s under a. On the offset range $0 \le i \le N - (t+1+L)$, define

$$\begin{aligned}Y_i &= H_i(h)\mathbf{1}\{s_{i+t+1} = s,\ a_{i+t+1} = a,\ (o_{i+t+1}, s_{i+t+2}) = (o^\star, s^\star)\},\\ Z_i &= H_i(h)\mathbf{1}\{a_{i+t+1} = a\} - Y_i.\end{aligned}$$

The first indicator counts occurrences of h in state s followed by action a and its uncorrupted joint outcome. The second counts the remaining occurrences followed by a: either the state is not s, or the joint outcome deviates from the uncorrupted outcome at s. The offset range leaves a complete length-L test after the one-step extension of h. Using the filtration in eq. (27), write the two centered sums as

$$M_U = \sum_{i=0}^{N-(t+1+L)} \left(U_i - \mathbb{E}[U_i \mid \mathcal{F}_{i+t}]\right), \qquad U \in \{Y, Z\}. \tag{34}$$

The dependence of these quantities on the fixed triple (h, s, a) is suppressed.

Theorem 5 (Simultaneous rows and counts)**.** *With probability at least $1 - \delta/2$, the centered count bounds in eq.* (30) *hold simultaneously for every action-observation history h with $|h| \le K$ and every action sequence $g \in \mathcal{A}^L$. In particular, on this event,*

$$\left|N(h,g) - \frac{N_h}{G}\right| \le b_h, \qquad \frac{N_h}{G} \ge 64L\Lambda \quad\Longrightarrow\quad \frac{N_h}{2G} \le N(h,g) \le \frac{3N_h}{2G}. \tag{35}$$

If $n_{\min} \ge 64L\Lambda$, then every row passing the history frequency filter satisfies

$$\begin{aligned}\left\|\hat{\mathcal{M}}^{K,L}_{h,:} - \overline{\mathbf{Forw}}^K_{h,:}\mathbf{Back}^L\right\|_2 &\le 2\sqrt{G}\,\eta(n_{\min}),\\ \max_{g\in\mathcal{A}^L}\left\|\hat{\mathcal{M}}^{K,L}_{h,:}|_g - \left(\overline{\mathbf{Forw}}^K_{h,:}\mathbf{Back}^L\right)|_g\right\|_\infty &\le \zeta(n_{\min}),\end{aligned} \tag{36}$$

where η and ζ are defined in eq. (21)*. The same event also gives*

$$|M_Y|,\ |M_Z| \le 2\sqrt{\frac{N_h\Lambda}{n}} + 3\Lambda \tag{37}$$

for the two sums in eq. (34)*, simultaneously for every history h with $|h| < K$, state s, and action a.*

Proof. There are $\sum_{j=0}^{K}(nm)^j$ action-observation histories of length at most K. For each history and each of the G action sequences, section D.3.4 uses L residue classes for each of the d observation sequence counts and the action-block count. These contribute $GL(d+1)\sum_{j=0}^{K}(nm)^j$ scalar martingales.

For the one-step counts, both Y_i and Z_i are indicators bounded above by $H_i(h)\mathbf{1}\{a_{i+t+1} = a\}$. Their conditional expectations given $\mathcal{F}_{i+t}$ are therefore at most $H_i(h)/n$. Each centered increment in eq. (34) has absolute value at most one, conditional variance at most $H_i(h)/n$, and is revealed by $\mathcal{F}_{i+t+1}$. Successive offsets thus form a martingale without a residue-class split, and the total predictable variance is at most

$$\frac{1}{n}\sum_{i=0}^{N-(t+1+L)} H_i(h) \le \frac{N_h}{n} \le N.$$

Empty sums are zero. Lemma 2 gives eq. (37) for these two sums, which contribute $2rn\sum_{j=0}^{K-1}(nm)^j$ scalar martingales across all triples (h, s, a). The two families together contain at most F_0 sums, as defined in eq. (20). Each has failure probability at most $2J_N e^{-\Lambda}$, so a union bound gives total failure probability at most

$$2F_0 J_N e^{-\Lambda} = \frac{\delta}{2}.$$

On the complementary event, recombining the residue classes gives eq. (30) simultaneously, and the denominator argument in section D.3.4 gives eq. (35). The union ranges over the entire deterministic history alphabet, so it includes whichever histories and one-step extensions the algorithm later selects.

For a row passing the frequency filter with $n_{\min} \ge 64L\Lambda$, decompose each block error as

$$\begin{aligned}\hat{\mathcal{M}}^{K,L}_{h,:}|_g - \left(\overline{\mathbf{Forw}}^{K}_{h,:}\mathbf{Back}^{L}\right)|_g = &\left[\hat{\mathcal{M}}^{K,L}_{h,:}|_g - \left(\overline{\mathbf{Forw}}^{K}_{h,:}\widetilde{\mathbf{Back}}^{L}\right)|_g\right] \\ &+ \left[\overline{\mathbf{Forw}}^{K}_{h,:}\left(\widetilde{\mathbf{Back}}^{L} - \mathbf{Back}^{L}\right)\right]|_g.\end{aligned}$$

The first term satisfies the sampling bounds in eq. (33), and lemma 3 bounds the second term for every realization of the finite-record weights. The triangle inequality yields

$$\begin{aligned}\left\|\hat{\mathcal{M}}^{K,L}_{h,:}|_g - \left(\overline{\mathbf{Forw}}^{K}_{h,:}\mathbf{Back}^{L}\right)|_g\right\|_\infty &\le 8\sqrt{\frac{L\Lambda}{n_{\min}}} + \eta_\beta = \zeta(n_{\min}), \\ \left\|\hat{\mathcal{M}}^{K,L}_{h,:}|_g - \left(\overline{\mathbf{Forw}}^{K}_{h,:}\mathbf{Back}^{L}\right)|_g\right\|_2 &\le 4(\sqrt{d}+1)\sqrt{\frac{L\Lambda}{n_{\min}}} + 2\eta_\beta = 2\eta(n_{\min}).\end{aligned}$$

Taking the maximum over action blocks proves the coordinate bound in eq. (36). Summing the squared Euclidean errors over the G disjoint blocks gives

$$\left\|\hat{\mathcal{M}}^{K,L}_{h,:} - \overline{\mathbf{Forw}}^{K}_{h,:}\mathbf{Back}^{L}\right\|_2^2 \le G\left(2\eta(n_{\min})\right)^2,$$

which proves the full-row bound. □

The square-root terms in these bounds arise from sampling, while η_β accounts for deviations during the test windows. The empirical distribution of states reached after the counted sequences equal to each history is retained in $\overline{\mathbf{Forw}}^{K}$; its perturbation for the supplied anchor histories is bounded separately in lemma 13.

D.4 Anchor Sequences and Backward Matrix Geometry

D.4.1 Identifying Histories and Anchor Supply

A *homing action sequence* is a fixed sequence of actions for which the resulting observation sequence uniquely determines the terminal state in the uncorrupted model, regardless of the starting state. Different observation sequences may identify different terminal states. We first obtain such an action sequence from a classical homing construction, then use connectivity to obtain an anchor sequence for each desired state.

Lemma 4 (Anchor supply)**.** *Under the connectivity and state-separation assumptions in section D.2.1, there exists a homing action sequence of length at most $(r-1)L$. Any homing action sequence of length L_{hom} gives an anchor sequence of length at most $L_{\mathrm{hom}} + D$ for every state. Consequently, the minimum anchor-length bound K_* is finite and satisfies*

$$K_* \le (r-1)L + D. \tag{38}$$

Proof. The homing construction in the proof of Kohavi & Jha (2009, Theorem 13.1, p. 437) concatenates at most $r-1$ action sequences, each separating a pair of distinct states. Our state-separation assumption supplies such a sequence of length L for every pair: distinct rows of $\mathbf{Back}^L$ differ in some action block, whose point masses specify different observation sequences. Applying that construction therefore gives a fixed homing action sequence of length at most $(r-1)L$.

Now take any homing action sequence of length L_{hom} and fix one feasible observation sequence under it. All uncorrupted executions producing this action-observation history end at the same state, say s_0. For a desired state s, strong connectivity supplies an action sequence of length at most D taking s_0 to s. Append those actions and their uniquely determined observations to the homing history. The resulting history is feasible, and every uncorrupted execution producing it first reaches s_0 after the homing prefix and then reaches s after the appended actions. It is therefore an anchor sequence for s, of length at most $L_{\mathrm{hom}} + D$. Applying this argument to every state proves eq. (38). When $r = 1$, the empty action sequence is already homing and $D = 0$, so the empty history anchors the only state and $K_* = 0$. □

The bound shows that $K \ge (r-1)L + D + 1$ is sufficient for the history-length requirement in eq. (17). The anchors already fixed in eq. (19) are chosen with lengths at most K_*, so their one-step extensions have length at most $K_* + 1 \le K$ and identify the corresponding uncorrupted successors. The later finite-count bounds establish how often these fixed histories occur in the corrupted trajectory.

D.4.2 Mixture Norms and Hull Separation

We now use the overlap bound in eq. (18) to quantify the geometry of the uncorrupted backward rows and their mixtures. Recall that every backward row has norm $\sqrt{G}$, and $u = 1 - \widehat{c} \ge 1/G > 0$.

Lemma 5 (Mixture norm bound)**.** *For any probability column vector $b \in \mathbb{R}^r$ indexed by $s \in \mathcal{S}$,*

$$1 - \frac{\|b^\top \cdot \mathbf{Back}^L\|_2^2}{G} \ge u\big(1 - \|b\|_2^2\big). \tag{39}$$

In particular, $\|b^\top \cdot \mathbf{Back}^L\|_2 = \sqrt{G}$ if and only if b is a point mass. For $\rho \in [0,1]$, if $\max_{s\in\mathcal{S}}[b]_s \le 1-\rho$, then

$$\frac{\|b^\top \cdot \mathbf{Back}^L\|_2^2}{G} \le 1 - u\rho. \tag{40}$$

Proof. Expanding the squared norm and separating the diagonal terms gives

$$\begin{aligned}\frac{\|b^\top \cdot \mathbf{Back}^L\|_2^2}{G} &= \sum_{s\in\mathcal{S}} [b]_s^2 + \sum_{s\ne t} [b]_s [b]_t c_{st} \\ &\le \|b\|_2^2 + \widehat{c}\big(1 - \|b\|_2^2\big) \\ &= 1 - u\big(1 - \|b\|_2^2\big),\end{aligned}$$

where the inequality uses $c_{st} \le \widehat{c}$, nonnegative weights, and $\sum_s [b]_s = 1$. This proves eq. (39). Since $u > 0$, a mixture with norm $\sqrt{G}$ must have $\|b\|_2^2 = 1$, which for a probability vector means that b is a point mass. Conversely, a point mass selects a single backward row, whose norm is $\sqrt{G}$. Finally,

$$\|b\|_2^2 = \sum_s [b]_s^2 \le \Big(\max_s [b]_s\Big) \sum_s [b]_s = \max_s [b]_s,$$

so $\max_s [b]_s \le 1 - \rho$ and eq. (39) imply eq. (40). □

For a history h with $N_h > 0$, these bounds apply with $b = (\overline{\mathbf{Forw}}^K_{h,:})^\top$, relating the norm of its reference row to the empirical distribution of states reached after the counted sequences equal to h. The next bound concerns the convex hull used by the stopping test in eq. (12), which includes the origin.

Lemma 6 (Hull separation of backward-matrix rows)**.** *Define*

$$\gamma = \min_{s \in \mathcal{S}} \operatorname{dist}_2 \Big(\mathbf{Back}^L_{s,:}, \operatorname{conv}\big(\{0\} \cup \{\mathbf{Back}^L_{t,:} : t \ne s\}\big) \Big). \tag{41}$$

Then

$$\gamma \ge \sqrt{G}\, u \ge \frac{1}{\sqrt{G}}.$$

Proof. Fix a state s and any row vector z in the convex hull appearing in eq. (41). Because the hull includes the origin, we can write

$$z = \sum_{t \ne s} [\lambda]_t \mathbf{Back}^L_{t,:}, \qquad [\lambda]_t \ge 0, \qquad \sum_{t \ne s} [\lambda]_t \le 1.$$

The overlap bound gives

$$\begin{aligned} \big\langle \mathbf{Back}^L_{s,:}, \mathbf{Back}^L_{s,:} - z \big\rangle &= G - G \sum_{t \ne s} [\lambda]_t c_{st} \\ &\ge G \left(1 - \widehat{c} \sum_{t \ne s} [\lambda]_t \right) \\ &\ge G(1 - \widehat{c}) = Gu. \end{aligned}$$

By Cauchy–Schwarz and $\|\mathbf{Back}^L_{s,:}\|_2 = \sqrt{G}$,

$$\sqrt{G}\, \|\mathbf{Back}^L_{s,:} - z\|_2 \ge \big\langle \mathbf{Back}^L_{s,:}, \mathbf{Back}^L_{s,:} - z \big\rangle \ge Gu.$$

Taking the infimum over z and the minimum over s gives $\gamma \ge \sqrt{G}\, u$, and $u \ge 1/G$ gives the second inequality. When $r = 1$, the hull is $\{0\}$ and $u = 1$, so the same argument gives $\gamma = \sqrt{G}$. □

Thus a missing state's backward row remains at distance at least $1/\sqrt{G}$ from the convex hull of the origin and any subset of the other backward rows. Together with the empirical-row error bounds, this separation will rule out premature stopping in theorem 7.

D.4.3 The Vertex Margin after the Khatri–Rao Product

For a positive integer C, recall the normalized C-fold Khatri–Rao product Φ_C from eq. (11). For an uncorrupted state s, $\Phi_C(\mathbf{Back}^L_{s,:})$ is the corresponding counterfactual backward row scaled by $G^{-C/2}$ and has unit norm. For a probability column vector b, the two expressions

$$\Phi_C(b^\top \cdot \mathbf{Back}^L) \qquad \text{and} \qquad \sum_{s \in \mathcal{S}} [b]_s \Phi_C(\mathbf{Back}^L_{s,:})$$

can differ, although they agree when b is a point mass. The algorithm takes the product of each empirical row and projects it against the span of products of selected rounded rows.

For a proper subset $P \subsetneq \mathcal{S}$, let R_P be the orthogonal projector onto the complement of

$$\operatorname{span}\{\Phi_C(\mathbf{Back}^L_{t,:})^\top : t \in P\},$$

with $R_\varnothing = I$. Define

$$\Delta_C = \min_{P \subsetneq \mathcal{S}} \min_{s \notin P} \left\| R_P \Phi_C(\mathbf{Back}^L_{s,:})^\top \right\|_2. \tag{42}$$

This is the smallest residual of an unselected state's product after projection against products of any subset of the other state rows. When the selected rounded rows equal the backward rows indexed by P, R_P is the projector used by the algorithm. Since the product rows have unit norm and orthogonal projection cannot increase norm, $0 \le \Delta_C \le 1$.

Let $Q_C \in \mathbb{R}^{r \times r}$ be the Gram matrix of these uncorrupted product rows. By eq. (13), its entries are

$$(Q_C)_{s,t} = \left\langle \Phi_C(\mathbf{Back}^L_{s,:}), \Phi_C(\mathbf{Back}^L_{t,:}) \right\rangle = \begin{cases} 1, & s = t, \\ c^C_{st}, & s \neq t. \end{cases}$$

Lemma 7 (Spectral lower bound)**.** *The margin in eq.* (42) *satisfies*

$$\Delta_C^2 \ge \lambda_{\min}(Q_C) \ge 1 - (r-1)\widehat{c}^C, \tag{43}$$

where $\lambda_{\min}(Q_C)$ is the smallest eigenvalue of Q_C. In particular,

$$(r-1)\widehat{c}^C \le \frac{1}{4} \quad \Longrightarrow \quad \Delta_C \ge \frac{\sqrt{3}}{2}. \tag{44}$$

Proof. Fix $P \subsetneq \mathcal{S}$ and $s \notin P$. The difference between $\Phi_C(\mathbf{Back}^L_{s,:})^\top$ and any vector in the selected span can be written as

$$\sum_{t \in \mathcal{S}} [x]_t \Phi_C(\mathbf{Back}^L_{t,:})^\top, \qquad [x]_s = 1, \qquad [x]_t = 0 \text{ for } t \notin P \cup \{s\},$$

where the column vector $x \in \mathbb{R}^r$ has unrestricted coordinates on P. The coefficient vector satisfies $\|x\|_2^2 \ge 1$ because $[x]_s = 1$. Since Q_C is a Gram matrix, it is positive semidefinite, and hence

$$\begin{aligned} \left\| \sum_{t \in \mathcal{S}} [x]_t \Phi_C(\mathbf{Back}^L_{t,:})^\top \right\|_2^2 &= x^\top Q_C x \\ &\ge \lambda_{\min}(Q_C)\|x\|_2^2 \\ &\ge \lambda_{\min}(Q_C). \end{aligned}$$

Taking the infimum over the coordinates of x on P gives the squared distance to that span, namely $\|R_P \Phi_C(\mathbf{Back}^L_{s,:})^\top\|_2^2$. Minimizing over P and $s \notin P$ proves the first inequality in eq. (43).

For the second inequality, take any column vector $x \in \mathbb{R}^r$. The unit diagonal and the off-diagonal bounds $0 \le c^C_{st} \le \widehat{c}^C$ give

$$\begin{aligned} x^\top Q_C x &= \|x\|_2^2 + \sum_{s \neq t} c^C_{st} [x]_s [x]_t \\ &\ge \|x\|_2^2 - \widehat{c}^C \sum_{s \neq t} |[x]_s [x]_t| \\ &\ge \left[1 - (r-1)\widehat{c}^C\right] \|x\|_2^2. \end{aligned}$$

The last step follows by summing $2|[x]_s[x]_t| \le [x]_s^2 + [x]_t^2$ over ordered pairs $s \neq t$ and dividing by two, giving a coefficient of $r-1$ for each squared coordinate. Taking the minimum Rayleigh quotient proves the second inequality, and eq. (44) follows. □

For $r \ge 2$, a sufficient number of counterfactual tests for eq. (44) is any positive integer C satisfying

$$C \ge \frac{\log(4(r-1))}{u}, \tag{45}$$

because $\widehat{c}^C = (1-u)^C \le e^{-uC}$. Such a finite choice exists since $u > 0$. If $r = 1$ or $\widehat{c} = 0$, then $Q_C = I$ and $\Delta_C = 1$ for every positive integer C. The recovery guarantee will compare the empirical-row perturbations with the actual margin Δ_C, while eq. (45) provides a sufficient bound on C for this margin to be at least $\sqrt{3}/2$.

D.5 Row Perturbations and Algorithm Termination

D.5.1 Rounding and Tensor Perturbations

The following arguments are deterministic once the empirical-row error bounds hold. For each history h under consideration, assume that h passes the frequency filter and satisfies

$$\begin{aligned} \big\|\hat{\mathcal{M}}^{K,L}_{h,:} - \overline{\mathbf{Forw}}^{K}_{h,:}\mathbf{Back}^{L}\big\|_2 &\le 2\sqrt{G}\,\eta, \\ \max_{g\in\mathcal{A}^L}\big\|\hat{\mathcal{M}}^{K,L}_{h,:}|_g - (\overline{\mathbf{Forw}}^{K}_{h,:}\mathbf{Back}^{L})|_g\big\|_\infty &\le \zeta. \end{aligned} \tag{46}$$

Here $\eta, \zeta \ge 0$ are common error bounds for the histories under consideration. With $n_{\min} \ge 64L\Lambda$, theorem 5 supplies these bounds simultaneously for every history passing the frequency filter, using $\eta = \eta(n_{\min})$ and $\zeta = \zeta(n_{\min})$ from eq. (21). The coordinate bound gives exact rounding, while the Euclidean bound gives the perturbation estimates used in projection and stopping.

Lemma 8 (Local rounding)**.** *Let h pass the frequency filter and satisfy the coordinate bound in eq.* (46)*. If, for some state s and $\rho \ge 0$,*

$$\overline{\mathbf{Forw}}^{K}_{h,s} \ge 1-\rho, \qquad \rho+\zeta < \frac{1}{2},$$

then the rounding rule in eq. (15) *returns $\mathcal{R}(h) = \mathbf{Back}^{L}_{s,:}$.*

Proof. We fix an action sequence $g \in \mathcal{A}^L$, and let $y \in \mathcal{O}^L$ be its unique observation sequence from s in the uncorrupted model, so that $\mathbf{Back}^{L}_{s,(g,y)} = 1$. All summands with $t \ne s$ are nonnegative, so

$$\Big[\overline{\mathbf{Forw}}^{K}_{h,:}\mathbf{Back}^{L}\Big]_{(g,y)} = \sum_{t\in\mathcal{S}} \overline{\mathbf{Forw}}^{K}_{h,t}\mathbf{Back}^{L}_{t,(g,y)} \ge \overline{\mathbf{Forw}}^{K}_{h,s} \ge 1-\rho.$$

The coordinate error bound therefore implies

$$\hat{\mathcal{M}}^{K,L}_{h,(g,y)} \ge 1-\rho-\zeta > \frac{1}{2}.$$

Because h passes the frequency filter, its empirical g-block is nonnegative and sums to one, so every other entry $\hat{\mathcal{M}}^{K,L}_{h,(g,y')}$ with $y' \ne y$ is at most $1 - \hat{\mathcal{M}}^{K,L}_{h,(g,y)} < 1/2$. Thus y uniquely maximizes $\hat{\mathcal{M}}^{K,L}_{h,(g,y')}$ over observation sequences $y' \in \mathcal{O}^L$, and $\widehat{y}_h(g) = y$. We apply this argument to every action sequence g, so rounding returns $\mathcal{R}(h) = \mathbf{Back}^{L}_{s,:}$. □

Lemma 9 (Khatri–Rao Lipschitz bound)**.** *For $G > 0$ and a positive integer C, define*

$$\Phi_C(z) = \left(\frac{z}{\sqrt{G}}\right)^{\odot C}.$$

If x and y are row vectors in the same finite-dimensional real Euclidean space with $\|x\|_2, \|y\|_2 \le \sqrt{G}$, then

$$\|\Phi_C(x) - \Phi_C(y)\|_2 \le \frac{C}{\sqrt{G}}\|x-y\|_2.$$

Proof. First, we normalize both x and y, yielding vectors $u = x/\sqrt{G}$ and $v = y/\sqrt{G}$ so that the norms $\|u\|_2$ and $\|v\|_2$ are at most 1. We then unpack a telescoping sum inside the norm of $\|\Phi_C(x) - \Phi_C(y)\|_2$:

$$
\begin{aligned}
\|\Phi_C(x) - \Phi_C(y)\|_2 &= \|u^{\odot C} - v^{\odot C}\|_2 \\
&= \left\| \sum_{j=0}^{C-1} u^{\odot j} \odot (u - v) \odot v^{\odot (C-1-j)} \right\|_2 \\
&\leq \sum_{j=0}^{C-1} \left\| u^{\odot j} \odot (u - v) \odot v^{\odot (C-1-j)} \right\|_2 \\
&= \sum_{j=0}^{C-1} \|u\|_2^j \, \|u - v\|_2 \, \|v\|_2^{C-1-j} \\
&\leq C\|u - v\|_2 = \frac{C}{\sqrt{G}} \|x - y\|_2.
\end{aligned}
$$

The first equality substitutes in the normalized vectors and the second unpacks the telescoping sum. The third step applies the triangle inequality, and the fourth step applies the identity that $\|a \odot b\|_2 = \|a\|_2 \, \|b\|_2$. The last inequality bounds the norm factors by one, since $\|u\|_2, \|v\|_2 \leq 1$, leaving C copies of $\|u - v\|_2$. □

For any history satisfying eq. (46), both $\hat{\mathcal{M}}^{K,L}_{h,:}$ and $\overline{\mathbf{Forw}}^{K}_{h,:}\mathbf{Back}^{L}$ have norm at most $\sqrt{G}$, as shown in sections D.1.1 and D.3.1. Applying lemma 9 to these two rows gives

$$
\begin{aligned}
&\left\|\Phi_C(\hat{\mathcal{M}}^{K,L}_{h,:}) - \Phi_C(\overline{\mathbf{Forw}}^{K}_{h,:}\mathbf{Back}^{L})\right\|_2 \\
&\quad \leq \frac{C}{\sqrt{G}} \left\|\hat{\mathcal{M}}^{K,L}_{h,:} - \overline{\mathbf{Forw}}^{K}_{h,:}\mathbf{Back}^{L}\right\|_2 \leq 2C\eta.
\end{aligned} \tag{47}
$$

Recall the fixed uncorrupted anchor history $h^\star_s$ for each state s from section D.2.2. On a record for which all these histories belong to E, let $\xi \in [0, 1]$ bound, for each anchor history, the fraction of its counted occurrences that reach a state other than the one it identifies:

$$
\max_s \left(1 - \overline{\mathbf{Forw}}^{K}_{h^\star_s, s}\right) \leq \xi.
$$

Here $1 - \overline{\mathbf{Forw}}^{K}_{h^\star_s, s}$ is the fraction of counted occurrences of $h^\star_s$ that end outside s. This fraction can be positive under corruption even though the history identifies s in the uncorrupted model.

The designated anchors are used to establish the recovery guarantee; the algorithm need not know or select them. We reserve ρ for the analogous bound on the histories selected by the algorithm. The selection guarantee will require $0 \leq \xi \leq \rho < 1/2$.

For the following bounds, let h be any history passing the frequency filter and satisfying eq. (46), and fix a state s such that $1 - \overline{\mathbf{Forw}}^{K}_{h,s} \leq \xi$. In particular, the displayed bound on ξ gives this inequality for each designated history $h = h^\star_s$.

Importantly, since each row of $\overline{\mathbf{Forw}}^{K}$ still forms a well-defined probability distribution, we have:

$$
\begin{aligned}
&\overline{\mathbf{Forw}}^{K}_{h,:}\mathbf{Back}^{L} - \mathbf{Back}^{L}_{s,:} \\
&\quad = \sum_t \overline{\mathbf{Forw}}^{K}_{h,t}\mathbf{Back}^{L}_{t,:} - \sum_t \overline{\mathbf{Forw}}^{K}_{h,t}\mathbf{Back}^{L}_{s,:} \\
&\quad = \sum_t \overline{\mathbf{Forw}}^{K}_{h,t} \left(\mathbf{Back}^{L}_{t,:} - \mathbf{Back}^{L}_{s,:}\right) \\
&\quad = \sum_{t \neq s} \overline{\mathbf{Forw}}^{K}_{h,t} \left(\mathbf{Back}^{L}_{t,:} - \mathbf{Back}^{L}_{s,:}\right).
\end{aligned}
$$

The total weight on states other than s is at most ξ:

$$
\sum_{t \neq s} \overline{\mathbf{Forw}}^{K}_{h,t} = 1 - \overline{\mathbf{Forw}}^{K}_{h,s} \leq \xi.
$$

Each uncorrupted backward row has one entry equal to one in each of its G action blocks, giving the following norm and distance bounds:

$$\|\mathbf{Back}^L_{t,:}\|_2 = \sqrt{G} \qquad (t = 1, \dots, r),$$
$$\|\mathbf{Back}^L_{t,:} - \mathbf{Back}^L_{s,:}\|_2 \le \|\mathbf{Back}^L_{t,:}\|_2 + \|\mathbf{Back}^L_{s,:}\|_2 = 2\sqrt{G}.$$

Applying the triangle inequality to the weighted differences, then using nonnegative weights and the mass bound outside s, gives

$$\begin{aligned}
&\Big\|\overline{\mathbf{Forw}}^K_{h,:}\mathbf{Back}^L - \mathbf{Back}^L_{s,:}\Big\|_2 \\
&\quad\le \sum_{t\neq s} \overline{\mathbf{Forw}}^K_{h,t} \Big\|\mathbf{Back}^L_{t,:} - \mathbf{Back}^L_{s,:}\Big\|_2 \\
&\quad\le 2\sqrt{G}\sum_{t\neq s} \overline{\mathbf{Forw}}^K_{h,t} \\
&\quad\le 2\sqrt{G}\,\xi.
\end{aligned}$$

Combining this bound with the empirical-row error yields

$$\begin{aligned}
&\big\|\hat{\mathcal{M}}^{K,L}_{h,:} - \mathbf{Back}^L_{s,:}\big\|_2 \\
&\quad\le \big\|\hat{\mathcal{M}}^{K,L}_{h,:} - \overline{\mathbf{Forw}}^K_{h,:}\mathbf{Back}^L\big\|_2 + \big\|\overline{\mathbf{Forw}}^K_{h,:}\mathbf{Back}^L - \mathbf{Back}^L_{s,:}\big\|_2 \\
&\quad\le 2\sqrt{G}\,(\xi+\eta).
\end{aligned} \tag{48}$$

Finally, $\|\hat{\mathcal{M}}^{K,L}_{h,:}\|_2 \le \sqrt{G}$ and $\|\mathbf{Back}^L_{s,:}\|_2 = \sqrt{G}$, so another application of lemma 9 gives

$$\begin{aligned}
&\big\|\Phi_C(\hat{\mathcal{M}}^{K,L}_{h,:}) - \Phi_C(\mathbf{Back}^L_{s,:})\big\|_2 \\
&\quad\le \frac{C}{\sqrt{G}}\big\|\hat{\mathcal{M}}^{K,L}_{h,:} - \mathbf{Back}^L_{s,:}\big\|_2 \le 2C(\xi+\eta).
\end{aligned} \tag{49}$$

These estimates apply to any history and state satisfying the stated bounds, including histories other than the designated anchors. Here ξ bounds the fraction of counted action-observation sequences equal to h that reach a state other than s, while η bounds the empirical row's error relative to its own finite-record reference mixture.

D.5.2 Near-Duplicate Residuals

We next bound the projection residual of a mixture concentrated near a state whose backward row has already been selected. For the reference mixture, the bound is independent of the number C of counterfactual tests. The empirical-row error is then added using eq. (47).

Lemma 10 (Near-duplicate residual)**.** *Let $b \in \mathbb{R}^r$ be a probability column vector, and suppose $[b]_s = 1-\theta$ for some state s and $0 \le \theta < 1$. For every positive integer C,*

$$\operatorname{dist}_2\Big(\Phi_C(b^\top \cdot \mathbf{Back}^L)^\top, \operatorname{span}\{\Phi_C(\mathbf{Back}^L_{s,:})^\top\}\Big)^2 \le \frac{\theta}{1-\theta}. \tag{50}$$

Consequently, let h pass the frequency filter and satisfy the Euclidean bound in eq. (46)*. If $\overline{\mathbf{Forw}}^K_{h,s} \ge 1-\rho$ for $0 \le \rho < 1$ and $s \in P \subsetneq \mathcal{S}$, then*

$$\big\|R_P\Phi_C(\hat{\mathcal{M}}^{K,L}_{h,:})^\top\big\|_2 \le \sqrt{\frac{\rho}{1-\rho}} + 2C\eta. \tag{51}$$

Proof. If $\theta = 0$, then b is a point mass at s, so the distance in eq. (50) is zero. For $0 < \theta < 1$, define the row vector

$$w = \frac{1}{\theta\sqrt{G}}\sum_{t\neq s}[b]_t\mathbf{Back}^L_{t,:}.$$

The weights $[b]_t/\theta$ for $t \neq s$ are nonnegative and sum to one, so w is a convex combination of normalized backward rows and $\|w\|_2 \leq 1$. The normalized mixture therefore decomposes as

$$\frac{b^\top \cdot \mathbf{Back}^L}{\sqrt{G}} = (1-\theta)\frac{\mathbf{Back}^L_{s,:}}{\sqrt{G}} + \theta w.$$

Set

$$\begin{aligned} c &= \left\langle \frac{\mathbf{Back}^L_{s,:}}{\sqrt{G}}, w \right\rangle, \\ a &= \frac{\langle b^\top \cdot \mathbf{Back}^L, \mathbf{Back}^L_{s,:} \rangle}{G} = 1 - \theta + \theta c, \\ q &= \frac{\|b^\top \cdot \mathbf{Back}^L\|_2^2}{G} = (1-\theta)^2 + 2\theta(1-\theta)c + \theta^2 \|w\|_2^2. \end{aligned}$$

Here c is the inner product of the normalized state row with the remaining mixture, a is its inner product with the full normalized mixture, and q is the full mixture's squared norm after normalization. Nonnegativity and the norm bounds give $0 \leq c \leq 1$, and Cauchy–Schwarz gives $0 \leq a^2 \leq q \leq 1$.

Since $\Phi_C(\mathbf{Back}^L_{s,:})$ has unit norm, the squared distance to its span is the squared norm of the mixture product minus its squared inner product with that state product. Using eq. (13), this gives

$$\mathrm{dist}_2\Big(\Phi_C(b^\top \cdot \mathbf{Back}^L)^\top, \mathrm{span}\{\Phi_C(\mathbf{Back}^L_{s,:})^\top\}\Big)^2 = q^C - a^{2C}.$$

If $c = 1$, then

$$\left\| w - \frac{\mathbf{Back}^L_{s,:}}{\sqrt{G}} \right\|_2^2 = \|w\|_2^2 + 1 - 2c \leq 0,$$

so $w = \mathbf{Back}^L_{s,:}/\sqrt{G}$ and the distance is again zero. For $c < 1$, expanding the definitions of q and a gives

$$\begin{aligned} q - a^2 &= \theta^2(\|w\|_2^2 - c^2) \leq \theta^2(1 - c^2), \\ 1 - q &= 2\theta(1-\theta)(1-c) + \theta^2(1 - \|w\|_2^2) \\ &\geq 2\theta(1-\theta)(1-c) > 0. \end{aligned}$$

In particular $q < 1$, so factoring the difference of powers and bounding the geometric sum yields

$$\begin{aligned} q^C - a^{2C} &= (q - a^2) \sum_{j=0}^{C-1} q^{C-1-j} a^{2j} \\ &\leq (q - a^2) \sum_{k=0}^{C-1} q^k \\ &\leq \frac{q - a^2}{1 - q} \\ &\leq \frac{\theta(1+c)}{2(1-\theta)} \leq \frac{\theta}{1-\theta}. \end{aligned}$$

The first inequality uses $a^{2j} \leq 1$ and $q - a^2 \geq 0$ and the second bounds the finite sum by the convergent infinite geometric series. This proves eq. (50).

To obtain the residual bound for the empirical SDM row, take $b = (\overline{\mathbf{Forw}}^K_{h,:})^\top$ and $\theta = 1 - \overline{\mathbf{Forw}}^K_{h,s} \leq \rho$. Since $s \in P$, the span removed by R_P contains $\Phi_C(\mathbf{Back}^L_{s,:})^\top$, so the reference mixture's residual is at most its distance to that one-dimensional span. The function $\theta/(1-\theta)$ increases on $[0, 1)$, giving

$$\big\| R_P \Phi_C(\overline{\mathbf{Forw}}^K_{h,:} \mathbf{Back}^L)^\top \big\|_2 \leq \sqrt{\frac{\rho}{1-\rho}}.$$

An orthogonal projector cannot increase Euclidean norm, so the triangle inequality and eq. (47) now give

$$\begin{aligned}&\left\|R_P\Phi_C(\hat{\mathcal{M}}^{K,L}_{h,:})^\top\right\|_2\\&\quad\le \left\|R_P\Phi_C(\overline{\mathbf{Forw}}^K_{h,:}\mathbf{Back}^L)^\top\right\|_2+\left\|\Phi_C(\hat{\mathcal{M}}^{K,L}_{h,:})-\Phi_C(\overline{\mathbf{Forw}}^K_{h,:}\mathbf{Back}^L)\right\|_2\\&\quad\le \sqrt{\frac{\rho}{1-\rho}}+2C\eta,\end{aligned}$$

which proves eq. (51). □

D.5.3 Distinct Selection and Exact Stopping

We now compare the projected norms of histories concentrated near missing states with the projected norms of the other candidates. The selection argument requires, for each state, a history in E with large enough mass on that state. The fixed anchor histories will supply these histories in the recovery guarantee; the following theorem applies to any histories satisfying its mass condition.

Theorem 6 (Distinct selection)**.** *Let $0 \le \xi \le \rho < 1/2$, and suppose every history $h \in E$ satisfies eq. (46) with common $\eta, \zeta \ge 0$. Suppose that, for every state s, there exists a history $h \in E$ such that $\overline{\mathbf{Forw}}^K_{h,s} \ge 1-\xi$. If $\rho+\zeta<1/2$ and*

$$\max\left\{(1-u\rho)^{C/2}, \sqrt{\frac{\rho}{1-\rho}}\right\}+2C\xi+4C\eta<\Delta_C, \tag{52}$$

then each of the first r selections, if made, chooses a history h with $\overline{\mathbf{Forw}}^K_{h,s}\ge 1-\rho$ for a distinct state s and rounds to $\mathbf{Back}^L_{s,:}$. These selections pass the duplicate-row and zero-residual checks.

Proof. We induct on the number of selected histories. Suppose their rounded rows are exactly $\mathbf{Back}^L_{t,:}$ for the states t in a proper subset $P \subsetneq \mathcal{S}$, and for each selected history, at least a fraction $1-\rho$ of its counted occurrences reach its corresponding state. The algorithm's projector is then R_P. This holds initially with $P=\varnothing$ and $R_\varnothing = I$.

For any missing state $s \notin P$, choose a history $h \in E$ with $\overline{\mathbf{Forw}}^K_{h,s}\ge 1-\xi$. This history remains available for selection. Indeed, if it had already been selected for a state $t \in P$, its probability vector would satisfy

$$\overline{\mathbf{Forw}}^K_{h,s}+\overline{\mathbf{Forw}}^K_{h,t}\ge 2-\xi-\rho>1,$$

contradicting the sum of its weights being one. By eq. (49) and the definition of Δ_C, its projected norm satisfies

$$\begin{aligned}&\left\|R_P\Phi_C(\hat{\mathcal{M}}^{K,L}_{h,:})^\top\right\|_2\\&\quad\ge \left\|R_P\Phi_C(\mathbf{Back}^L_{s,:})^\top\right\|_2-\left\|\Phi_C(\hat{\mathcal{M}}^{K,L}_{h,:})-\Phi_C(\mathbf{Back}^L_{s,:})\right\|_2\\&\quad\ge \Delta_C-2C(\xi+\eta).\end{aligned}\tag{53}$$

First consider a candidate history h with $\max_s \overline{\mathbf{Forw}}^K_{h,s}<1-\rho$. By lemma 5, the kernel identity, and eq. (47),

$$\begin{aligned}\left\|R_P\Phi_C(\hat{\mathcal{M}}^{K,L}_{h,:})^\top\right\|_2&\le \left\|\Phi_C(\overline{\mathbf{Forw}}^K_{h,:}\mathbf{Back}^L)\right\|_2+2C\eta\\&=\left(\frac{\|\overline{\mathbf{Forw}}^K_{h,:}\mathbf{Back}^L\|_2^2}{G}\right)^{C/2}+2C\eta\\&\le (1-u\rho)^{C/2}+2C\eta.\end{aligned}$$

Next consider a candidate history h with $\overline{\mathbf{Forw}}^K_{h,t}\ge 1-\rho$ for a state $t\in P$. The near-duplicate bound in lemma 10 gives

$$\left\|R_P\Phi_C(\hat{\mathcal{M}}^{K,L}_{h,:})^\top\right\|_2\le\sqrt{\frac{\rho}{1-\rho}}+2C\eta.$$

Condition eq. (52) makes both of these upper bounds strictly smaller than the available missing-state history's lower bound in eq. (53). Consequently, for any maximizing candidate h, at least a fraction $1-\rho$ of the counted action-observation sequences equal to h reach a missing state $s \notin P$. Such a state is unique because $1-\rho > 1/2$.

By lemma 8 and $\rho+\zeta < 1/2$, the chosen history rounds to $\mathbf{Back}^L_{s,:}$. Our state separation assumption ensures this row is distinct from all previously selected rows. Moreover, eq. (52) implies $\Delta_C > 0$, so the rounded row's projected product satisfies

$$\left\| R_P \Phi_C(\mathbf{Back}^L_{s,:})^\top \right\|_2 \geq \Delta_C > 0.$$

Thus, neither the duplicate-row check nor the zero-residual check rejects the selection. The projection update removes the newly selected direction, establishing the induction for the enlarged selected span. □

Theorem 7 (Exact stopping)**.** *Under the hypotheses of theorem 6, if the stopping tolerance satisfies*

$$2\sqrt{G}\,\eta \leq \tau < \gamma - 2\sqrt{G}\,(\xi+\eta), \tag{54}$$

then the selection stage stops after exactly r selections and proceeds to matrix construction with $\widehat{r} = r$. Its selected rounded rows are precisely the r uncorrupted backward rows, in some order. For the algorithm's choice $\tau = 1/(2\sqrt{G})$, a sufficient additional condition is

$$\xi + \eta < \frac{1}{4G}.$$

Proof. The available history for each state ensures that E is nonempty. Before r states have been selected, theorem 6 identifies the rounded rows with a proper subset $P \subsetneq \mathcal{S}$ of the uncorrupted backward rows. Choose a missing state $s \notin P$ and a history $h \in E$ with $\overline{\mathbf{Forw}}^K_{h,s} \geq 1-\xi$. The current stopping hull satisfies

$$\mathcal{C}_F = \operatorname{conv}\big(\{0\} \cup \{\mathbf{Back}^L_{t,:} : t \in P\}\big) \subseteq \operatorname{conv}\big(\{0\} \cup \{\mathbf{Back}^L_{t,:} : t \neq s\}\big).$$

Therefore, the distance from $\mathbf{Back}^L_{s,:}$ to $\mathcal{C}_F$ is at least γ. The triangle inequality for distance to a set and eq. (48) give

$$\begin{aligned}\operatorname{dist}_2(\hat{\mathcal{M}}^{K,L}_{h,:}, \mathcal{C}_F) &\geq \operatorname{dist}_2(\mathbf{Back}^L_{s,:}, \mathcal{C}_F) - \|\hat{\mathcal{M}}^{K,L}_{h,:} - \mathbf{Back}^L_{s,:}\|_2 \\ &\geq \gamma - 2\sqrt{G}\,(\xi+\eta) > \tau,\end{aligned}$$

where the strict inequality uses eq. (54). The stopping test thus fails before r selections, including at the initial empty list where $\mathcal{C}_F = \{0\}$. The preceding proof also shows that a missing state's history remains available and that each subsequent selection passes the duplicate-row and zero-residual checks. Hence, the algorithm reaches r selections.

At that point, every history $h \in E$ has reference row

$$\overline{\mathbf{Forw}}^K_{h,:}\mathbf{Back}^L = \sum_{s \in \mathcal{S}} \overline{\mathbf{Forw}}^K_{h,s}\mathbf{Back}^L_{s,:} \in \mathcal{C}_F,$$

because the weights are nonnegative and sum to one and all backward rows have been selected. Its empirical row consequently satisfies

$$\operatorname{dist}_2(\hat{\mathcal{M}}^{K,L}_{h,:}, \mathcal{C}_F) \leq \left\|\hat{\mathcal{M}}^{K,L}_{h,:} - \overline{\mathbf{Forw}}^K_{h,:}\mathbf{Back}^L\right\|_2 \leq 2\sqrt{G}\,\eta \leq \tau.$$

This holds for every history in the fixed set E, so eq. (12) succeeds with a nonempty list of r rounded rows.

For $\tau = 1/(2\sqrt{G})$, the condition $\xi + \eta < 1/(4G)$ gives

$$2\sqrt{G}\,\eta \leq 2\sqrt{G}\,(\xi+\eta) < \tau.$$

By lemma 6, $\gamma \geq 1/\sqrt{G}$, so it also gives

$$\gamma - 2\sqrt{G}\,(\xi+\eta) > \frac{1}{\sqrt{G}} - \frac{1}{2\sqrt{G}} = \tau.$$

Thus, both inequalities in eq. (54) hold. □

If a positive known lower bound $\gamma_0 \le \gamma$ is available, the same argument permits $\tau = \gamma_0/2$ with $\xi + \eta < \gamma_0/(4\sqrt{G})$. The stopping proof uses the convex hull of the origin and the selected rounded rows, which equal the uncorrupted backward rows once all states have been selected. This hull contains every reference mixture.

D.6 Perturbation of Supplied Anchors and History Visit Counts

D.6.1 Uncorrupted Replay and Finite Counts

We next bound how often the histories in the fixed family $\mathcal{H}$, consisting of the anchor histories and their one-step extensions under the uncorrupted model, occur in the recorded trajectory. An occurrence of a nonempty history h is a sequence of consecutive recorded action-observation pairs that equals h. Call an occurrence of h *uncorrupted* if, at every step in that recorded sequence, the observation and successor state agree with the uncorrupted model at the actual current state and action. This condition concerns the history window; the subsequent test may contain deviations. Recall $W = D + K$, $\alpha_\beta = ((1-\beta)/n)^W$, and $\mu = \lfloor (N-L)/W \rfloor \alpha_\beta$ from section D.2.2. Here D is the diameter of the uncorrupted transition graph, K is the history-length bound, $n = |\mathcal{A}|$ is the number of actions, and β bounds the probability of a joint observation and successor-state deviation at each step. The record contains N action-observation pairs, and L is the test length, so $\lfloor (N-L)/W \rfloor$ counts full blocks that leave room for a subsequent test. The quantity α_β is the lower bound on the probability of an uncorrupted occurrence of a given history in each block established below, and μ is this bound multiplied by the number of usable blocks. Since $K \ge 1$, we have $W \ge 1$.

Lemma 11 (Uncorrupted recurrence in a block)**.** *Fix a nonempty history h of length at most K that can occur in the uncorrupted model. Let q be a fixed integer with $0 \le q \le N - W$, representing the number of trajectory steps before the block $q+1, \dots, q+W$. Given all preceding actions, observations, and latent states, including the current state s_{q+1}, the conditional probability that h has an uncorrupted occurrence wholly within the block is at least α_β. Its completion time lies strictly after q and no later than $q + W$.*

Proof. Choose a state from which the uncorrupted model produces h when its action sequence is executed. Strong connectivity supplies an action sequence of length $0 \le \ell \le D$ from the actual state s_{q+1} to the chosen state capable of producing h. Follow this sequence and then the actions in h. At each step, conditional on the preceding actions and outcomes, the probability of the specified action and its uncorrupted joint outcome is at least $(1-\beta)/n$ by uniform action sampling and eq. (16). Sequential conditioning therefore bounds the probability of this entire execution below by

$$\left(\frac{1-\beta}{n}\right)^{\ell+|h|} \ge \left(\frac{1-\beta}{n}\right)^{D+K} = \alpha_\beta.$$

On this event, the recorded action-observation pairs at steps $q+\ell+1, \dots, q+\ell+|h|$ form the sequence h, with the last pair at step $q+\ell+|h| \in (q, q+W]$. The reaching sequence may depend on s_{q+1}, but the probability bound holds uniformly over that state and the preceding trajectory. □

For the empty history, no recorded action-observation pair is required. For the block ending at step $q + W$, use the empty sequence immediately after that step, before the next action is taken. This sequence contains no outcomes that could be corrupted, so it is uncorrupted with probability one.

Recall that N_h counts observed occurrences of h with a full subsequent length-L test window, whether or not their history windows are uncorrupted.

Lemma 12 (Finite counts)**.** *With probability at least $1 - |\mathcal{H}|e^{-\mu/8}$, every $h \in \mathcal{H}$ has at least $\mu/2$ uncorrupted occurrences among those counted by N_h. In particular, $N_h \ge \mu/2$ simultaneously for every $h \in \mathcal{H}$.*

Proof. Partition the first $\lfloor (N-L)/W \rfloor W$ trajectory steps into consecutive blocks of length W. Index the blocks by integers $b = 0, \dots, \lfloor (N-L)/W \rfloor - 1$, so $q = bW$ for block b, which consists of steps $bW + 1, \dots, (b+1)W$. Write $\sum_b$ for the sum over all these blocks, and take empty sums to be zero. Fix $h \in \mathcal{H}$, and let Z_b equal one if at least one uncorrupted occurrence of h lies wholly

within block b, and zero otherwise. For the empty history, set $Z_b = 1$ using the block's final time. Recall the trajectory filtration

$$\mathcal{F}_j = \sigma\big(s_1, (a_i, o_i, s_{i+1})_{i=1}^{j}\big), \qquad 0 \le j \le N,$$

where $\mathcal{F}_0 = \sigma(s_1)$ and $\mathcal{F}_j$ contains the actions, observations, and latent states through step j, including the state s_{j+1} before the next action. Thus $\mathcal{F}_{bW}$ describes the information at the start of block b, and Z_b is $\mathcal{F}_{(b+1)W}$-measurable. Every history in $\mathcal{H}$ is feasible in the uncorrupted model and has length at most K, so lemma 11 gives

$$\mathbb{P}\{Z_b = 1 \mid \mathcal{F}_{bW}\} \ge \alpha_\beta.$$

Consequently, for every $\lambda > 0$,

$$\begin{aligned}\mathbb{E}[e^{-\lambda Z_b} \mid \mathcal{F}_{bW}] &= 1 - (1 - e^{-\lambda})\mathbb{P}\{Z_b = 1 \mid \mathcal{F}_{bW}\} \\ &\le \exp\{-\alpha_\beta(1 - e^{-\lambda})\}.\end{aligned}$$

For a fixed $\lambda > 0$, define

$$M_k = \exp\left\{-\lambda \sum_{b=0}^{k-1} Z_b + k\alpha_\beta(1 - e^{-\lambda})\right\}, \qquad 0 \le k \le \left\lfloor \frac{N-L}{W} \right\rfloor.$$

The process (M_k) is adapted to $(\mathcal{F}_{kW})$, and the preceding bound gives

$$\begin{aligned}\mathbb{E}[M_{k+1} \mid \mathcal{F}_{kW}] &= M_k \exp\{\alpha_\beta(1 - e^{-\lambda})\}\mathbb{E}[e^{-\lambda Z_k} \mid \mathcal{F}_{kW}] \\ &\le M_k.\end{aligned}$$

Thus (M_k) is a nonnegative supermartingale with $M_0 = 1$, so $\mathbb{E}[M_k] \le 1$ for every k. The block indicators need not be independent. Set $\lambda = \log 2$ and use $\mu = \lfloor (N-L)/W \rfloor \alpha_\beta$. Markov's inequality applied to the terminal value of this supermartingale gives

$$\begin{aligned}\mathbb{P}\left\{\sum_b Z_b < \frac{\mu}{2}\right\} &= \mathbb{P}\left\{M_{\lfloor (N-L)/W \rfloor} > \exp\left(\frac{1 - \log 2}{2}\mu\right)\right\} \\ &\le \exp\left\{-\frac{1 - \log 2}{2}\mu\right\} \mathbb{E}[M_{\lfloor (N-L)/W \rfloor}] \\ &\le \exp\left\{-\frac{1 - \log 2}{2}\mu\right\} \le e^{-\mu/8}.\end{aligned}$$

It remains to check that the blocks with $Z_b = 1$ give distinct occurrences of h counted by the estimator. Write $t = |h|$, and choose one supplied completion time $e_b \in (bW, (b+1)W]$ from each successful block. These times are distinct, and each is at most $\lfloor (N-L)/W \rfloor W \le N - L$. For nonempty h, the recorded action-observation sequence equal to h begins at step $i_b + 1$, where $i_b = e_b - t$ is the number of preceding steps. Because this sequence lies wholly within block b and ends by step $N - L$, we have

$$0 \le bW \le i_b \le N - (t + L).$$

The same inequalities hold for the empty history with $t = 0$ and $i_b = e_b = (b+1)W$. By construction, $(ao)_{i_b+1:i_b+t} = h$, so each successful block contributes a distinct uncorrupted occurrence of h to N_h with a complete subsequent test window. A union bound over the fixed family $\mathcal{H}$ proves the claim. □

D.6.2 Global Deviations and Anchor Perturbation

For an anchor history h identifying a state s in the uncorrupted model, we now bound the fraction of counted action-observation sequences equal to h for which the state reached after the sequence differs from s. For each trajectory step $q = 1, \ldots, N$, let E_q equal one if the observation o_q or successor state s_{q+1} differs from the deterministic outcome prescribed by the uncorrupted model at (s_q, a_q), and zero otherwise. Recall that $\beta = 2\epsilon < 1$ bounds the conditional probability of such a

joint deviation for every current state and action, by the observation and transition error bounds in eq. (16). Write

$$D_q = \sum_{j=1}^{q} E_j, \qquad 0 \le q \le N,$$

for the number of joint deviations in the first q steps, with $D_0 = 0$. Recall the deviation bound introduced in section D.2.2:

$$\Lambda_c = \log \frac{4}{\delta}, \qquad B_N = \begin{cases} 0, & \beta = 0, \\ 2\beta N + \Lambda_c, & \beta > 0. \end{cases}$$

Lemma 13 (Anchor perturbation)**.** *With probability at least $1 - \delta/4$, the total number of joint deviations satisfies $D_N \le B_N$. On this event, simultaneously for every anchor history h identifying a state s in the uncorrupted model, with $|h| \le K$ and $N_h > 0$,*

$$1 - \overline{\mathbf{Forw}}^{K}_{h,s} \le \frac{KB_N}{N_h}.$$

If $\mu > 0$, then on the intersection with the occurrence-count event in lemma 12, every history $h \in \mathcal{H}$ and the state s it identifies satisfy

$$1 - \overline{\mathbf{Forw}}^{K}_{h,s} \le \kappa_N = \frac{2KB_N}{\mu}.$$

Proof. Use the trajectory filtration $\mathcal{F}_q$ recalled in the preceding proof, so $\mathcal{F}_{q-1}$ contains the actions, observations, and latent states before step q, including its starting state s_q. The joint deviation bound in eq. (16) holds for every action, so averaging over the next action gives

$$\mathbb{E}[E_q \mid \mathcal{F}_{q-1}] \le \beta.$$

Since E_q takes values in $\{0, 1\}$,

$$\mathbb{E}[e^{E_q} \mid \mathcal{F}_{q-1}] = 1 + (e-1)\mathbb{E}[E_q \mid \mathcal{F}_{q-1}] \le e^{(e-1)\beta}.$$

Thus the process

$$M_q = \exp\{D_q - (e-1)\beta q\}, \qquad 0 \le q \le N,$$

is a nonnegative supermartingale with respect to $(\mathcal{F}_q)$, because

$$\begin{aligned} \mathbb{E}[M_q \mid \mathcal{F}_{q-1}] &= M_{q-1} e^{-(e-1)\beta} \mathbb{E}[e^{E_q} \mid \mathcal{F}_{q-1}] \\ &\le M_{q-1}. \end{aligned}$$

Its initial value is $M_0 = 1$, so $\mathbb{E}[M_N] \le 1$. For $\beta > 0$, Markov's inequality at the terminal step gives

$$\begin{aligned} \mathbb{P}\{D_N > 2\beta N + \Lambda_c\} &= \mathbb{P}\{M_N > e^{(3-e)\beta N + \Lambda_c}\} \\ &\le e^{-(3-e)\beta N - \Lambda_c} \le e^{-\Lambda_c} = \frac{\delta}{4}, \end{aligned}$$

where $e < 3$. When $\beta = 0$, each $E_q = 0$ almost surely, so $D_N = 0 = B_N$ almost surely.

Fix an anchor history h identifying s, and first suppose its length $t = |h|$ is positive. Consider recorded steps $i+1, \dots, i+t$ whose action-observation sequence equals h, with $0 \le i \le N-(t+L)$. The state reached after this sequence is s_{i+t+1}. If none of these t steps has a joint deviation, the sequence is an uncorrupted execution of h, so the state reached must be s, regardless of the starting state. Thus $s_{i+t+1} \ne s$ requires a deviation during these t steps, giving

$$\mathbf{1}\{(ao)_{i+1:i+t} = h,\ s_{i+t+1} \ne s\} \le \sum_{q=i+1}^{i+t} E_q.$$

A deviation at step q lies in a length-t window starting at $i+1$ only if $q-t \le i \le q-1$, so it belongs to at most t such windows. Using the finite-record state frequencies from eq. (23) and summing over the counted windows therefore gives

$$N_h\big(1 - \overline{\mathbf{Forw}}^{K}_{h,s}\big) \le \sum_{i=0}^{N-(t+L)} \sum_{q=i+1}^{i+t} E_q \le tD_N \le KD_N.$$

If h is empty and identifies a state regardless of its starting state, the model has only that one state, so $1 - \overline{\mathbf{Forw}}^K_{h,s} = 0$. Dividing by $N_h > 0$ proves the first perturbation bound whenever $D_N \le B_N$. This argument holds for every identifying history on the same deviation event, so no additional union bound over histories is needed.

Every history in $\mathcal{H}$ identifies a state in the uncorrupted model and has length at most K. On the event in lemma 12, $N_h \ge \mu/2$ for all of these histories. When $\mu > 0$, substitution gives

$$1 - \overline{\mathbf{Forw}}^K_{h,s} \le \frac{KB_N}{N_h} \le \frac{2KB_N}{\mu} = \kappa_N.$$

In particular, $\kappa_N = 0$ when $\beta = 0$ and $\mu > 0$. □

D.7 Observation and Transition Recovery

Theorem 6 ensures that, for each selected history h, at least a fraction $1 - \rho$ of the counted action-observation sequences equal to h reach its corresponding state. Such a history need not identify that state in every uncorrupted execution, so the matrix-recovery argument must apply to the one-step extensions of any history the algorithm might select.

D.7.1 The Finite-Record Parent–Child Identity

Let h be a history of length $t = |h| < K$, and fix an action $a \in \mathcal{A}$ and observation $o \in \mathcal{O}$. Call h the parent history and $h' = h \cdot (a, o)$ its child, obtained by appending the pair (a, o). Use i to index recorded sequences beginning at step $i + 1$, so i counts the preceding trajectory steps. The parent sequence uses steps $i+1, \dots, i+t$, and its one-step extension also includes the pair at step $i+t+1$. To leave a full length-L test after this child, we need

$$0 \le i \le N - (t + 1 + L).$$

Recall that $H_i(h) = \mathbf{1}\{(ao)_{i+1:i+t} = h\}$ indicates whether the recorded action-observation sequence at these t steps equals h. The recorded sequence of length $t + 1$ equals h' exactly when its first t pairs form h and its final pair is (a, o), giving

$$\begin{aligned} H_i(h') &= H_i(h)\mathbf{1}\{a_{i+t+1} = a,\ o_{i+t+1} = o\}, \\ N_{h'} &= \sum_{i=0}^{N-(t+1+L)} H_i(h)\mathbf{1}\{a_{i+t+1} = a,\ o_{i+t+1} = o\}. \end{aligned} \tag{55}$$

When the recorded sequence is h', the state reached after its parent sequence h is s_{i+t+1}, and the state reached after the full child sequence h' is s_{i+t+2}.

The full parent count N_h uses the larger range $0 \le i \le N - (t + L)$. Appending one pair shifts the test one step later, so the last starting position available to the parent no longer leaves L subsequent steps for the child. When the parent range is nonempty, the only excluded index is $i = N - (t + L)$, which contributes either zero or one to N_h. Consequently, the number of starting positions where the recorded sequence equals h and leaves room for the extension and its subsequent test satisfies

$$N_h - 1 \le \sum_{i=0}^{N-(t+1+L)} H_i(h) \le N_h. \tag{56}$$

Only positions where the recorded sequence h is followed by the pair (a, o) contribute to the child count $N_{h'}$. Both index ranges have the same lower limit, zero; appending the pair changes only the upper limit. All sums with a negative upper limit are zero. For an empty parent, $t = 0$ and $H_i(h) = 1$, and the child consists of the single pair (a, o), so the same identities apply.

D.7.2 Child State-Distribution Perturbation and Matrix Recovery

We now bound the distribution of states reached after the child sequences used to reconstruct the transition matrices. Recall that $\beta = 2\epsilon < 1$ bounds the conditional probability of a joint observation

and successor-state deviation, $n = |\mathcal{A}|$, and Λ is the concentration quantity in eq. (20). For a parent count $v > 0$, define the sampling term

$$z(v) = 2\sqrt{\frac{n\Lambda}{v}} + \frac{3n\Lambda}{v}.$$

For a given state s and a fixed bound $0 \le \rho < 1$ on the fraction of counted parent sequences reaching a different state, define

$$\rho_{\text{child}}(v) = \frac{\rho + \beta + z(v)}{1 - 1/v}, \qquad v > 1.$$

The numerator combines the parent perturbation, joint deviations during the additional action, and sampling error. The denominator accounts for the possible loss of one parent sequence when requiring a full test after the child.

Lemma 14 (Child state-distribution perturbation)**.** *On the simultaneous event in theorem 5, let h be a history with $|h| < K$ and $v = N_h > 1$, and suppose $\overline{\mathbf{Forw}}^K_{h,s} \ge 1 - \rho$ for a state s and $0 \le \rho < 1$. For any action a, let $o^\star$ and $s^\star$ be its uncorrupted observation and successor from s, and set $h' = h \cdot (a, o^\star)$. If $\rho_{\text{child}}(v) < 1$, then $N_{h'} > 0$ and*

$$1 - \overline{\mathbf{Forw}}^K_{h',s^\star} \le \rho_{\text{child}}(v).$$

If $\rho_{\text{child}}(v) \ge 1$, the same bound holds whenever $N_{h'} > 0$. These conclusions hold simultaneously for all histories, states, and actions satisfying the hypotheses, including histories selected using the recorded data.

Proof. Write $t = |h|$ and use the range $0 \le i \le N - (t + 1 + L)$, which leaves room for the child and a subsequent length-L test. All sums over i in this proof use this range. Recall the indicators from section D.3.5:

$$Y_i = H_i(h)\mathbf{1}\{s_{i+t+1} = s,\ a_{i+t+1} = a,\ (o_{i+t+1}, s_{i+t+2}) = (o^\star, s^\star)\},$$
$$Z_i = H_i(h)\mathbf{1}\{a_{i+t+1} = a\} - Y_i.$$

Here $H_i(h) = 1$ means that the recorded action-observation sequence at steps $i+1, \dots, i+t$ equals h. The indicator Y_i also requires that this sequence reach s and that the next action be a with its uncorrupted observation and successor. The indicator Z_i counts the remaining sequences equal to h followed by action a: either the state reached after h differs from s, or the additional action has a different joint outcome. Let $Y = \sum_i Y_i$ and $Z = \sum_i Z_i$.

The filtration $\mathcal{F}_{i+t}$ in eq. (27) contains the recorded pairs and latent states through step $i+t$, including the state before the additional action. The simultaneous deviation bound in eq. (37) gives

$$\left| Y - \sum_i \mathbb{E}[Y_i \mid \mathcal{F}_{i+t}] \right|, \quad \left| Z - \sum_i \mathbb{E}[Z_i \mid \mathcal{F}_{i+t}] \right| \le 2\sqrt{\frac{v\Lambda}{n}} + 3\Lambda = \frac{v}{n} z(v).$$

Thus the existing concentration event already supplies the sampling bounds needed here.

At least $(1 - \rho)v$ of the parent sequences counted by N_h reach s. Restricting their starting positions as in eq. (56) removes at most one of these sequences. Given a sequence reaching s, the conditional probability of action a followed by the uncorrupted joint outcome is at least $(1 - \beta)/n$. For the sequences counted by Z, at most ρv parent sequences reach a different state, and the probability of a joint deviation from s is at most β. The independent uniform actions and these bounds therefore give

$$\sum_i \mathbb{E}[Y_i \mid \mathcal{F}_{i+t}] \ge \frac{1 - \beta}{n}\big((1 - \rho)v - 1\big),$$
$$\sum_i \mathbb{E}[Z_i \mid \mathcal{F}_{i+t}] \le \frac{\rho + \beta}{n} v.$$

Combining with the sampling bounds yields

$$Y \ge \frac{v}{n}\big(1 - 1/v - \rho - \beta - z(v)\big) = \frac{v - 1}{n}\big(1 - \rho_{\text{child}}(v)\big),$$
$$Z \le \frac{v}{n}\big(\rho + \beta + z(v)\big) = \frac{v - 1}{n}\rho_{\text{child}}(v).$$

The lower bound uses $(1-\beta)((1-\rho)v-1) \geq v(1-\rho-\beta)-1$.

Let $N_{h',\neq s^\star}$ count the recorded sequences equal to h' that contribute to $N_{h'}$ and reach a state other than $s^\star$. The identity in eq. (55) shows that every sequence counted by Y contributes to $N_{h'}$ and reaches $s^\star$. Every sequence counted by $N_{h',\neq s^\star}$ contributes to Z, giving

$$N_{h',\neq s^\star} \leq Z, \qquad N_{h'} \geq Y + N_{h',\neq s^\star}.$$

If $\rho_{\text{child}}(v) < 1$, the lower bound on Y is positive, so $N_{h'} > 0$. For $y > 0$ and $w \geq 0$, the fraction $w/(y+w)$ increases with w and decreases with y. Using $Y \geq (v-1)(1-\rho_{\text{child}}(v))/n$ and $N_{h',\neq s^\star} \leq (v-1)\rho_{\text{child}}(v)/n$ therefore gives

$$1 - \overline{\mathbf{Forw}}^K_{h',s^\star} = \frac{N_{h',\neq s^\star}}{N_{h'}} \leq \frac{N_{h',\neq s^\star}}{Y + N_{h',\neq s^\star}} \leq \rho_{\text{child}}(v).$$

If $\rho_{\text{child}}(v) \geq 1$ and $N_{h'} > 0$, the fraction of sequences reaching a state other than $s^\star$ is at most one, so the same bound holds. The concentration event covers all parent histories, states, and actions simultaneously, which makes the conclusion valid for histories chosen using the data. □

Theorem 8 (Exact read-off)**.** *On the simultaneous event in theorem 5, suppose the selected rounded-row list consists of the r uncorrupted backward rows, each appearing exactly once. Let $h \in E$ be a selected history corresponding to state s, with*

$$\mathcal{R}(h) = \mathbf{Back}^L_{s,:}, \qquad \overline{\mathbf{Forw}}^K_{h,s} \geq 1-\rho, \qquad N_h > 1, \qquad 0 \leq \rho < 1.$$

Suppose every indicated child $h \cdot (a, o_h(a))$ satisfies the coordinate error bound in eq. (46) *with a common $\zeta \geq 0$. If*

$$\rho_{\text{child}}(N_h) + \zeta < \frac{1}{2},$$

then the matrix construction recovers the observation and transition rows corresponding to s correctly, and each of these child rows has a unique rounded-row match in the selected list. If these hypotheses hold for every selected history, the algorithm returns the uncorrupted observation and transition matrices under one common permutation of the states.

Proof. Fix an action a, and let $o^\star$ and $s^\star$ be its uncorrupted observation and successor from s. Since $\mathcal{R}(h) = \mathbf{Back}^L_{s,:}$, for each action sequence g the observation sequence $\widehat{y}_h(g)$ chosen by the rounding rule is the unique uncorrupted observation sequence produced from s under g. For every g beginning with a, its first observation is therefore $o^\star$. Thus $o_h(a) = o^\star$, and the child inspected by the algorithm is $h' = h \cdot (a, o^\star)$.

Because $h \in E$, this child has length at most K and passes the frequency filter. Lemma 14 and $\rho_{\text{child}}(N_h) + \zeta < 1/2$ give

$$\overline{\mathbf{Forw}}^K_{h',s^\star} \geq 1 - \rho_{\text{child}}(N_h), \qquad \mathcal{R}(h') = \mathbf{Back}^L_{s^\star,:},$$

where the rounding conclusion follows from lemma 8 using the child's coordinate error bound. The selected list contains $\mathbf{Back}^L_{s^\star,:}$ exactly once, so the algorithm finds a unique successor for action a. Together with $o_h(a) = o^\star$, this proves correctness of the observation and transition entries constructed from h.

If the hypotheses hold for every selected history, all child rows have unique matches and the matrix construction returns a model. Let $\sigma(i)$ be the position of $\mathbf{Back}^L_{s^i,:}$ in the selected list, for each true state s^i. This defines one permutation because the list consists of precisely these r rows, each appearing exactly once. The construction in section D.1.3 consequently satisfies

$$\widehat{T}^a_{\sigma(i),\sigma(j)} = T^a_{ij}, \qquad \widehat{O}^a_{\sigma(i),b} = O^a_{ib}$$

for every action a, state indices i, j, and observation index b. The observation columns retain their original alphabet labels. □

D.8 Composing the Main Result

We now combine the concentration, selection, occurrence, and child-perturbation bounds into a recovery guarantee. The frequency-filter threshold must be large enough for the empirical-row bounds and small enough for the fixed anchor histories and their children to pass the filter. For a given record length N, these two requirements give the integer interval

$$\lceil 64L\Lambda \rceil \le n_{\min} \le \left\lfloor \frac{\mu}{4G} \right\rfloor.$$

The theorem also requires a probability bound for the occurrences and bounds on the perturbations used in selection, stopping, and child rounding. Thus, a threshold in this interval must still satisfy the remaining conditions below.

Theorem 9 (General robust CfactSP guarantee)**.** *Under the model assumptions in section D.2.1, let $N \ge \max\{2, L+1\}$, $\delta \in (0,1)$, and $K \ge K_* + 1$. Run the procedure in section D.1 with positive integers $C, n_{\min}$ and stopping tolerance $\tau = 1/(2\sqrt{G})$. Use Λ, μ, κ_N from section D.2.2, and write $\eta = \eta(n_{\min})$ and $\zeta = \zeta(n_{\min})$ for the row-error quantities in eq.* (21)*. Suppose there exist $0 \le \xi \le \rho < 1/2$ satisfying the following conditions, with ρ_{child} from section D.7.2 defined using this same ρ:*

1. *Row concentration:*
$$n_{\min} \ge 64L\Lambda.$$
2. *Occurrence counts and their probability:*
$$\mu \ge \max\left\{4Gn_{\min},\ 8\log\frac{4r(n+1)}{\delta}\right\}.$$
3. *Perturbation of the supplied anchor histories:*
$$\kappa_N \le \xi.$$
4. *Distinct selection:*
$$\max\left\{(1-u\rho)^{C/2}, \sqrt{\frac{\rho}{1-\rho}}\right\} + 2C\xi + 4C\eta < \Delta_C.$$
5. *Exact stopping:*
$$\xi + \eta < \frac{1}{4G}.$$
6. *Child rounding:*
$$\rho_{\text{child}}(Gn_{\min}) + \zeta < \frac{1}{2}.$$

The first two conditions ensure $Gn_{\min} > 1$ and $\mu > 0$, so all these quantities are defined. Then, with probability at least $1-\delta$, the procedure selects exactly the r uncorrupted backward rows and returns $\widehat{r} = r$ without returning INCONCLUSIVE*. There is one permutation σ of the state indices such that, for every action a, state indices i, j, and observation index b,*

$$\widehat{T}^a_{\sigma(i),\sigma(j)} = T^a_{ij}, \qquad \widehat{O}^a_{\sigma(i),b} = O^a_{ib}.$$

The observation labels are unchanged. The guarantee assumes exact evaluation of the projections, hull distances, and rounding in the stated procedure.

Proof. Since $L \ge 1$ and $\Lambda > 1$, the row-concentration condition gives $Gn_{\min} > 1$. The definition of the child perturbation bound gives $\rho_{\text{child}}(Gn_{\min}) \ge \rho$. Consequently, the child-rounding condition also implies $\rho + \zeta < 1/2$, which will be used to round the parent histories.

Intersect the simultaneous count and row event of theorem 5, the global deviation event of lemma 13, and the occurrence event of lemma 12 for the fixed family $\mathcal{H}$ of anchors and their uncorrupted one-step extensions. These events have failure probabilities at most $\delta/2$, $\delta/4$, and $|\mathcal{H}|e^{-\mu/8}$, respectively. By eq. (19) and the occurrence condition,

$$|\mathcal{H}|e^{-\mu/8} \le r(n+1)e^{-\mu/8} \le \frac{\delta}{4}.$$

By a union bound, the count and row bounds, the global deviation bound, and the occurrence bounds for every history in $\mathcal{H}$ hold simultaneously with probability at least $1 - \delta/2 - \delta/4 - \delta/4 = 1 - \delta$. Suppose all these bounds hold. We show that the algorithm then recovers the uncorrupted model.

Every history $h \in \mathcal{H}$ has $N_h \geq \mu/2$. In particular,

$$\frac{N_h}{G} \geq \frac{\mu}{2G} \geq 2n_{\min} \geq 64L\Lambda.$$

The count implication in eq. (35) applies before imposing the frequency filter and gives

$$N(h, g) \geq \frac{N_h}{2G} \geq \frac{\mu}{4G} \geq n_{\min} \qquad \text{for every } g \in \mathcal{A}^L.$$

Thus, every supplied anchor and each of its uncorrupted one-step extensions passes the frequency filter. The simultaneous row bounds now apply to all of them.

For every state s, lemma 13 gives

$$1 - \overline{\mathbf{Forw}}^K_{h^\star_s, s} \leq \kappa_N \leq \xi \leq \rho.$$

Since $\rho + \zeta < 1/2$, lemma 8 gives $\mathcal{R}(h^\star_s) = \mathbf{Back}^L_{s,:}$. For each action a, the most frequent observation sequences chosen for this anchor therefore have the same first observation as the uncorrupted model whenever their action sequences begin with a. The children indicated by those first observations are precisely the extensions already included in $\mathcal{H}$, so they pass the frequency filter as shown above. Together with $|h^\star_s| \leq K_* < K$, these facts establish $h^\star_s \in E$. Hence E is nonempty and contains, for every state, a history with at least $1 - \xi$ of its finite-record state mass on that state.

The simultaneous row event supplies the deterministic error bounds for every history in E. The selection condition and $\rho + \zeta < 1/2$ therefore allow us to apply theorem 6. Each selection chooses a history with at least $1 - \rho$ of its mass on a previously unselected state and rounds to that state's backward row. The same theorem excludes duplicate rounded rows and zero projected norms of their Khatri–Rao products; in particular, the strict selection inequality ensures $\Delta_C > 0$. Before all states have been selected, a missing state's supplied history remains available. The stopping condition and theorem 7 prevent stopping before r selections, including with an empty selected list, and force stopping after the rth selection. Thus, none of the inconclusive cases in the selection stage occurs, and the selected list consists of the r uncorrupted backward rows, each exactly once.

Finally, every selected history h passes the frequency filter, so

$$N_h = \sum_{g \in \mathcal{A}^L} N(h, g) \geq G n_{\min} > 1.$$

For $v > 1$, the numerator $\rho + \beta + z(v)$ of $\rho_{\text{child}}(v)$ is nonnegative and decreases with v, while its positive denominator $1 - 1/v$ increases. Therefore,

$$\rho_{\text{child}}(N_h) + \zeta \leq \rho_{\text{child}}(G n_{\min}) + \zeta < \frac{1}{2}.$$

Membership in E supplies every indicated child within the history-length bound and passing the frequency filter, and the simultaneous row event supplies its coordinate error bound ζ. The hypotheses of theorem 8 consequently hold for every selected history. Every child has a unique rounded-row match, and the matrix construction returns the uncorrupted transition and observation matrices under one common state permutation. □

D.8.1 Choosing the Frequency Rate

The paper specifies the frequency filter through ν, with $n_{\min} = \lceil \nu N \rceil$. For a given record length, any positive integer threshold can be represented by taking $\nu = n_{\min}/N$. If ν is specified first, its rounded threshold must satisfy all the conditions of theorem 9, including

$$\lceil 64L\Lambda \rceil \leq \lceil \nu N \rceil \leq \left\lfloor \frac{\mu}{4G} \right\rfloor.$$

The lower bound ensures sufficiently accurate empirical rows, while the upper bound allows the supplied anchor histories and their children to pass the filter. The next construction gives a fixed positive choice of ν and explicit bounds on corruption and record length that satisfy the recovery conditions.

D.8.2 A Feasible Choice of Parameters

The following parameter choice satisfies the general recovery theorem with selected-history perturbation tolerance $\rho = 1/8$. Here $\alpha_0 = n^{-W}$ is the occurrence-probability bound at zero corruption, and $b_\star$ is an upper bound on the combined perturbation $\kappa_N + \eta(n_{\min})$.

Corollary 1 (A feasible frequency rate and corruption bound)**.** *Under the model assumptions in section D.2.1, fix a separating test length L and a history cap $K \geq K_* + 1$. Choose*

$$C = \left\lceil \frac{\max\{16\log 2, \log(4r)\}}{u} \right\rceil, \qquad b_\star = \min\left\{ \frac{1}{16C}, \frac{1}{8G} \right\}, \qquad \alpha_0 = n^{-W}.$$

Set the frequency rate and integer threshold to

$$\nu = \frac{\alpha_0}{32GW}, \qquad n_{\min} = \lceil \nu N \rceil.$$

Define the sample-size coefficient

$$N_{\text{scale}} = \frac{32W}{\alpha_0} \max\left\{ \frac{K}{b_\star}, \frac{64GL(\sqrt{d}+1)^2}{b_\star^2} \right\}.$$

This coefficient collects the requirements for row accuracy and the finite-record anchor perturbation. If

$$0 \leq \beta \leq \min\left\{ \frac{1}{2W}, \frac{b_\star \alpha_0}{64KW}, \frac{b_\star}{4L} \right\}, \tag{57}$$

then, for any $\delta \in (0,1)$ and integer record length satisfying

$$N \geq \max\left\{ 2(L+W),\ 2N_{\text{scale}} \log \frac{8N_{\text{scale}}F_0}{\delta} \right\}, \tag{58}$$

the procedure in section D.1, with stopping tolerance $\tau = 1/(2\sqrt{G})$, returns exactly r states and the uncorrupted transition and observation matrices under one common state permutation, with probability at least $1-\delta$. In particular, every fixed model satisfying the assumptions admits a positive corruption range and a fixed positive frequency rate for which recovery is guaranteed at all sufficiently large record lengths.

Proof. We verify the conditions of theorem 9 directly. The chosen number of counterfactual tests satisfies

$$C \geq \frac{16\log 2}{u} \geq 4, \qquad (r-1)\widehat{c}^{\,C} \leq re^{-uC} \leq \frac{1}{4}.$$

The spectral bound in lemma 7 therefore gives $\Delta_C \geq \sqrt{3}/2$.

First, the sample bound in eq. (58) implies $N \geq N_{\text{scale}}\Lambda$. Indeed, $J_N \leq N$ for integers $N \geq 2$, so $\Lambda \leq \log(4NF_0/\delta)$. The function $N/N_{\text{scale}} - \log(4NF_0/\delta)$ increases for $N > N_{\text{scale}}$ and is nonnegative at $N = 2N_{\text{scale}}\log(8N_{\text{scale}}F_0/\delta)$. For the latter claim, writing $x = \log(8N_{\text{scale}}F_0/\delta) > 1$ gives $\log(4NF_0/\delta) = x + \log x \leq 2x = N/N_{\text{scale}}$ at that value of N. Hence $N \geq N_{\text{scale}}\Lambda$ throughout the stated range. By the definitions of N_{scale} and ν,

$$\nu N \geq \frac{64L(\sqrt{d}+1)^2}{b_\star^2}\Lambda \geq 64L\Lambda > 1.$$

Thus $n_{\min} \geq 64L\Lambda$ and $n_{\min} \leq 2\nu N$.

The first corruption bound gives $(1-\beta)^W \geq 1 - W\beta \geq 1/2$, so $\alpha_\beta \geq \alpha_0/2$. Since $N \geq 2(L+W)$,

$$\mu = \left\lfloor \frac{N-L}{W} \right\rfloor \alpha_\beta \geq \frac{N\alpha_0}{4W} = 8G\nu N \geq 4Gn_{\min}.$$

Also, $N_{\text{scale}} \geq 32W/\alpha_0$ and $F_0 \geq 2rn \geq r(n+1)$ give

$$\mu \geq \frac{N\alpha_0}{4W} \geq 8\Lambda \geq 8\log\frac{4r(n+1)}{\delta}.$$

Both occurrence requirements therefore hold.

To bound the combined perturbation, use $B_N \le 2\beta N + \log(4/\delta)$, including when $\beta = 0$, and $\eta_\beta \le L\beta$ to obtain

$$\kappa_N + \eta(n_{\min}) \le \frac{16KW\beta}{\alpha_0} + \frac{8KW\log(4/\delta)}{N\alpha_0} + 2(\sqrt{d}+1)\sqrt{\frac{L\Lambda}{\nu N}} + L\beta.$$

The second and third bounds in eq. (57) bound the first and fourth terms by $b_\star/4$ each. The inequality $N \ge N_{\text{scale}}\Lambda$, together with $\Lambda \ge \log(4/\delta)$, bounds each of the other two terms by $b_\star/4$. Thus $\kappa_N + \eta(n_{\min}) \le b_\star$.

Set $\rho = 1/8$, $\xi = \kappa_N$, $\eta = \eta(n_{\min})$, and $\zeta = \zeta(n_{\min})$. Since $C \ge 4$ and $b_\star \le 1/(16C)$, we have $\xi \le 1/64 < \rho$. Moreover,

$$(1-u\rho)^{C/2} \le e^{-uC/16} \le \frac{1}{2}, \qquad \sqrt{\frac{\rho}{1-\rho}} = \frac{1}{\sqrt{7}} < \frac{1}{2}.$$

The bound $2C\xi + 4C\eta \le 4C(\xi+\eta) \le 4Cb_\star \le 1/4$ makes the left side of the selection condition at most $3/4 < \sqrt{3}/2 \le \Delta_C$. The stopping condition follows from $\xi + \eta \le b_\star \le 1/(8G) < 1/(4G)$.

It remains to check child rounding. The row-error formulas give $\zeta \le 2\eta$ because $d \ge 1$, and $\beta \le \eta_\beta \le \eta$ because $L \ge 1$. Consequently,

$$\beta \le \frac{1}{64}, \qquad \zeta \le \frac{1}{32}.$$

Using $G = n^L \ge n$, $n_{\min} \ge 64L\Lambda$, and $\Lambda > 1$ gives

$$Gn_{\min} \ge 64, \qquad \frac{n\Lambda}{Gn_{\min}} \le \frac{1}{64}, \qquad z(Gn_{\min}) \le \frac{1}{4} + \frac{3}{64} = \frac{19}{64}.$$

Therefore,

$$\rho_{\text{child}}(Gn_{\min}) + \zeta \le \frac{1/8 + 1/64 + 19/64}{1 - 1/64} + \frac{1}{32} = \frac{4}{9} + \frac{1}{32} = \frac{137}{288} < \frac{1}{2}.$$

All conditions of theorem 9 are now satisfied, giving the claimed recovery probability. All three upper bounds on β are strictly positive and independent of N and δ, and the chosen ν is fixed and positive. □

D.8.3 Worst-Case Bounds in Terms of the Number of States

We now derive the coarse bounds in theorem 3 from the feasible parameter choice. Assume that every pair of distinct states can be distinguished by some action sequence. To bound the required test length, group states according to whether they produce the same observations for every action sequence of length at most ℓ. At $\ell = 0$, all states belong to one group. Two states remain together at length $\ell + 1$ exactly when every action gives the same immediate observation and takes them to states in the same length-ℓ group. If a refinement makes no split, this characterization makes every subsequent refinement unchanged. The distinguishability assumption therefore forces a strict increase in the number of groups until all r states are separate, which occurs by length $r - 1$. Extending a shorter distinguishing sequence preserves its differing observation prefix, so we may take

$$L = \max\{1, r-1\}.$$

A shortest path between two states does not repeat a state, giving $D \le r - 1$. The anchor-length bound in lemma 4 then permits

$$K = (r-1)L + D + 1 \le r^2, \qquad W = D + K \le r^2.$$

These inequalities include $r = 1$, where $D = 0$ and $L = K = W = 1$. Since $u \ge 1/G$, the choices in corollary 1 satisfy

$$G = n^L, \qquad d = m^L, \qquad \alpha_0^{-1} = n^W, \qquad C = O\big(n^L \log(2r)\big), \qquad b_\star^{-1} = O\big(n^L \log(2r)\big).$$

The reciprocal of the permitted corruption bound in eq. (57) is

$$\max\left\{2W,\ \frac{64KW}{b_\star\alpha_0},\ \frac{4L}{b_\star}\right\} \le \max\{2,n\}^{O(r^2)}.$$

Here the largest action-alphabet exponent is $W + L = O(r^2)$, and the polynomial factors in r and $\log(2r)$ are absorbed into the same bound. Since $\beta = 2\epsilon$, there is a universal constant $c_1 > 0$ such that

$$\epsilon \le \max\{2,n\}^{-c_1 r^2}$$

implies the required corruption inequalities.

For the sample requirement, substitution into N_{scale} gives

$$N_{\mathrm{scale}} = O\big(Wn^W \max\left\{Kn^L \log(2r),\ Ln^{3L}m^L \log^2(2r)\right\}\big).$$

Since $K, W = O(r^2)$ and $L = O(r)$, this is at most $\max\{2,nm\}^{O(r^2)}$. The definition of F_0 has the same bound: its history sums contain at most $K + 1$ terms, each at most $(nm)^K$, and its remaining factors are polynomial in r, L, n and the test-alphabet sizes n^L, m^L. Consequently, the explicit sample requirement in eq. (58) is satisfied by

$$N \ge \max\{2,nm\}^{c_2 r^2} \log\frac{2}{\delta}$$

for a sufficiently large universal constant $c_2 > 0$. To see the confidence dependence, expand the logarithm there as $\log(8N_{\mathrm{scale}}F_0) + \log(1/\delta)$ and absorb the model-dependent logarithm into the preceding exponential bound. Thus only one factor of $\log(2/\delta)$ is required.

For $n \ge 2$, the corruption expression has the $n^{-O(r^2)}$ form stated informally in the main paper. The maximum with two also makes the bound meaningful when there is only one action. These bounds are deliberately loose: they replace the actual distinguishing length, anchor lengths, overlap, and occurrence probabilities by bounds using only r, n, m.