Multistep Inverse Is Not All You Need

Alexander Levine
alevine0@cs.utexas.edu
The University of Texas at Austin

Peter Stone
pstone@cs.utexas.edu
The University of Texas at Austin and Sony AI

Amy Zhang
amy.zhang@austin.utexas.edu
The University of Texas at Austin

Abstract

In real-world control settings, the observation space is often unnecessarily high-dimensional and subject to time-correlated noise. However, the *controllable* dynamics of the system are often far simpler than the dynamics of the raw observations. It is therefore desirable to learn an encoder to map the observation space to a simpler space of control-relevant variables. In this work, we consider the Ex-BMDP model, first proposed by Efroni et al. (2022b), which formalizes control problems where observations can be factorized into an action-dependent latent state which evolves deterministically, and action-independent time-correlated noise. Lamb et al. (2022) proposes the "AC-State" method for learning an encoder to extract a complete action-dependent latent state representation from the observations in such problems. AC-State is a *multistep-inverse* method, in that it uses the encoding of the the first and last state in a path to predict the *first* action in the path. However, we identify cases where AC-State will fail to learn a correct latent representation of the agent-controllable factor of the state. We therefore propose a new algorithm, ACDF, which combines multistep-inverse prediction with a latent forward model. ACDF is guaranteed to correctly infer an action-dependent latent state encoder for a large class of Ex-BMDP models. We demonstrate the effectiveness of ACDF on tabular Ex-BMDPs through numerical simulations; as well as high-dimensional environments using neural-network-based encoders. Code is available at `https://github.com/midi-lab/acdf`.

1 Introduction

In rich-observation decision-making domains, such as robotics, much of the information that the agent observes is irrelevant to any plausible control objective. To allow for efficient planning, it is therefore desirable to learn a compact latent state representation, containing only the information potentially relevant to planning. One approach to this problem is to learn a *control-endogenous latent representation* (Efroni et al., 2022b; Lamb et al., 2022). The intuition behind this approach is that the observations that an agent receives, such as images, may contain a large amount of irrelevant information (including time-correlated noise) which represents parts of the environment that the agent has no control over. By contrast, the agent-controllable dynamics of the system can in some cases be represented by a *small number of states with deterministic transitions.* This representation allows for efficient planning, and provides a view of the world model than can be directly interpreted by humans. The use of such representations has shown success, for example, in learning robotic manipulation tasks from images in a noisy environment (Lamb et al., 2022).

Efroni et al. (2022b) introduces the Ex-BMDP formalism to represent this kind of environment. An Ex-BMDP is a (reward-free) Markov Decision Process in which each *observed* state $x \in \mathcal{X}$ can be

factored into an agent-controllable *endogenous* state $s \in \mathcal{S}$, which follows a deterministic transition function, and an *exogenous* state $e \in \mathcal{E}$, which evolves stochastically, independently of actions. It is important to note that the observations x in this model are *not explicitly* segmented into factors: rather, in order to extract the controllable latent state s, we must *learn* an encoder.[1]

Numerous prior works have proposed methods for discovering latent representations useful for planning (Wang et al., 2022; Zhang et al., 2020; Pathak et al., 2017; Mazoure et al., 2020). We focus our attention on the multi-step inverse method (Lamb et al., 2022; Islam et al., 2023), which is compelling due to its explicit theoretical justification. In brief, consider states x_t and x_{t+k} visited by a policy. If an encoder ϕ is learned, such that $\phi(x_t)$ and $\phi(x_{t+k})$ provide sufficient information to predict the *first action* a_t on the path between x_t and x_{t+k}, then the learned representation $s = \phi(x)$ is claimed to be a complete endogenous latent state representation, providing the necessary and sufficient information to infer the latent dynamics.

This paper discusses cases where the multistep inverse method proposed by Lamb et al. (2022), known as AC-State, will fail to discover control-endogenous latent dynamics of an Ex-BMDP, and proposes a method that provably succeeds for a very general class of Ex-BMDP's. In particular, we identify two flaws with the AC-State method:

- The maximum length of the segment k between x_t and x_{t+k} required for multistep-inverse dynamics prediction in order to correctly learn the encoder can be much larger than claimed.
- If the dynamics are periodic, then multistep-inverse dynamics are insufficient for learning an appropriate encoder ϕ, regardless of k.

We then propose a modified method, which we call **ACDF**, which fixes these issues. We show that any encoder which minimizes our loss function (on infinite samples) is *guaranteed to be a control-endogenous latent representation*. Specifically, we:

- Give a corrected formulation of the number of steps of multistep-inverse dynamics prediction required to learn an Ex-BMDP.
- Propose to use a *latent forward dynamics loss*, to enforce that the learned endogenous states are in fact compatible with deterministic dynamics.

In addition to our theoretical claims, we show empirically that ACDF can produce a more accurate endogenous latent model in Ex-BMDPs exhibiting certain properties.

2 Background and Motivating Example

Here, we formally describe the Ex-BMDP model and AC-State algorithm, and provide a simple example in which AC-State fails:

2.1 Ex-BMDP Model

To formalize this notion of control-endogenous latent dynamics, consider a reward-free MDP with states $x \in \mathcal{X}$, discrete actions $a \in \mathcal{A}$, initial distribution $\mathcal{D}_0 \in \Delta(\mathcal{X})$, and transition function $x_{t+1} \sim \mathcal{T}(x|x_t, a_t)$. The MDP admits a *control endogenous latent representation* if its transition function can be decomposed as follows:

$$\begin{aligned} &x_{t+1} \sim \mathcal{Q}(x|s_{t+1}, e_{t+1}), \\ &s_{t+1} = T(s_t, a_t), \quad s_t = \phi(x_t), \\ &e_{t+1} \sim \mathcal{T}_e(e|e_t), \quad e_t = \phi_e(x_t), \end{aligned} \tag{1}$$

[1] Additionally, x can also have time-independent noise: the observation x is sampled from a distribution that depends only on s and e. It is assumed that the observation x contains enough information to fully specify s and e.

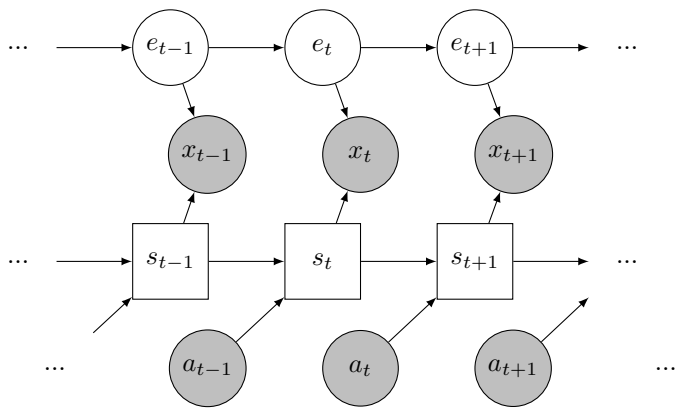


Figure 1: Probabilistic graphical model of the Ex-BMDP transition dynamics, as described in Section 2.1. Endogenous states s_t are shown as squares to indicate that they are *deterministic* functions of the previous endogenous states and actions. Observations x_t and actions a_t are shown in gray to indicate that they are observable. We do *not* show dependencies that may determine the actions a_t.

where:

- $s \in \mathcal{S}$ and $e \in \mathcal{E}$ are referred to as the *control-endogenous* and *control-exogenous* latent states, respectively. We assume that $\mathcal{S}$ is finite, and typically $|\mathcal{S}| \ll |\mathcal{X}|$.[2]
- The endogenous latent state s evolves according to a *deterministic* transition function T.
- The exogenous latent state e evolves according to a Markovian transition function $\mathcal{T}_e$ that *does not* depend on actions.
- We make the *block assumption* on the observation emission function $\mathcal{Q}$ (Du et al., 2019): that is, we assume that if $(s, e) \neq (s', e')$, then $\mathcal{Q}(x|s, e)$ and $\mathcal{Q}(x|s', e')$ have disjoint support. In other words, an observation $x \in \mathcal{X}$ corresponds to only a single pair (s, e).
- The encoders ϕ, ϕ_e are the (deterministic) inverses of $\mathcal{Q}$: that is, if $x \sim \mathcal{Q}(x|s, e)$, then $s = \phi(x)$, $e = \phi_e(x)$.

An MDP that admits such a representation is known as an Ex-BMDP. Note that an Ex-BMDP can have *multiple* valid factorizations into endogenous and exogenous states. (We discuss this fact further in Appendix C.) In this work, our objective is to learn the encoder ϕ that is the endogenous state encoder for some valid factorization of the Ex-BMDP: specifically, we will aim to return a *minimal-state* encoder: the number of endogenous states $|\mathcal{S}|$ should be as small as possible. We show a probabilistic graphical model of the Ex-BMDP transition dynamics in Figure 1.

2.2 AC-State

We now give more detail on the AC-State method as proposed by Lamb et al. (2022). In addition to the Ex-BMDP formulation, that work makes the following further "bounded diameter" assumption:

Assumption (Assumption 3.1 from Lamb et al. (2022))**.** "The length of the shortest path between any $z_1 \in \mathcal{S}$ to any $z_2 \in \mathcal{S}$ is bounded by D."

Lamb et al. (2022) then proposes to learn the endogenous encoder ϕ_θ by learning a *multi-step inverse dynamics model*. This model is a learned classifier $f(\phi_\theta(x_t), \phi_\theta(x_{t+k}); k)$ which takes the encoded endogenous latent states of two observations separated by $k \leq D$ time-steps, as well as k, and returns

[2]In our proofs, we also assume that $\mathcal{E}$ and $\mathcal{X}$ are finite, but this is a technical limitation of our theory: see discussion in Appendix B. In any case, the Ex-BMDP formulation is most useful when $|\mathcal{S}| \ll |\mathcal{X}|$.

a normalized distribution over *actions* in $\mathcal{A}$. This classifier is trained to predict the first action a_t taken on the trajectory from x_t to x_{t+k}. The classifier f is trained jointly with the encoder model. Then, in the theoretical treatment, the optimal encoder ϕ_{θ^*} is defined as the encoder which allows f to reach the minimum achievable value of this classification loss while also using the fewest number of distinct output states. (In practice, Lamb et al. (2022) uses a discrete information bottleneck and an associated loss term to minimize the output range, rather than learning multiple models.) Explicitly:

$$\begin{aligned}
&\mathcal{L}_{\text{AC-State}}(\phi_\theta) := \min_f \mathop{\mathbb{E}}_{k\sim\{1,\ldots,D\}} \mathop{\mathbb{E}}_{(x_t,a_t,x_{t+k})} -\log(f_{a_t}(\phi_\theta(x_t),\phi_\theta(x_{t+k});k)) \\
&\{\theta\}^* := \{\theta^{**} | \theta^{**} = \arg\min_\theta \mathcal{L}_{\text{AC-State}}(\phi_\theta)\} \\
&\theta^* := \arg\min_{\theta\in\{\theta\}^*} \|\text{Range}(\phi_\theta)\|
\end{aligned} \tag{2}$$

where $f_{a_t}(\cdot)$ represents the probability assigned by f to the action a_t. Given some assumptions about the behavioral policy (which are satisfied, for example, by a uniformly random policy), it is claimed that the ϕ_θ which minimizes this loss (assuming perfect function approximation and large numbers of samples) will produce a control-endogenous latent representation of the MDP. The transition function $T(s,a)$ can then be inferred after-the-fact by applying the learned ϕ to all observed states x and counting the transitions between the resulting latent states.

In practice, the diameter D of the endogenous latent dynamics of the Ex-BMDP is unknown *a priori*. Throughout this work, we will use D to represent the true diameter of the endogenous latent dynamics[3], and K to represent the number of steps actually used in practice. Based on the above assumption, AC-State is claimed by Lamb et al. (2022) to work as long as $K \geq D$.

However, we can demonstrate a simple case where AC-State *will not* recover a control-endogenous latent representation. In particular, consider the Ex-BMDP shown in Figure 2. If we focus our attention on the "correct" endogenous latent state dynamics shown in Figure 2-C, we can note that:

- We can't infer the action a_t from s_t and s_{t+k} if s_t is any state other than the state $s_t = \{a, j\}$, because for all other states in $\mathcal{S}_{\text{ACDF}}$, the actions L and R have the same effect.
- We can't infer a_t given s_t and s_{t+k} for any $k > 2$, because, once $\{a, j\}$ is visited a second time, a_t no longer has any impact on the current state.
- The only remaining case is that $s_t = \{a, j\}$ and $k \in \{1, 2\}$. In this case, if s_{t+k} is *either* $\{b, g\}$ *or* $\{c, h\}$, then we know that $a_t = L$. Similarly, if s_{t+k} is *either* $\{d, i\}$ *or* $\{e, j\}$, then we know that $a_t = R$.

Note that there are *no* cases where predicting a_t requires distinguishing between the states $\{b, g\}$ and $\{c, h\}$, or distinguishing between the states $\{d, i\}$ and $\{e, j\}$. Therefore the optimal minimal-state AC-State encoder, the ϕ_{θ^*} produced by Equation 2, will distinguish between only three "states": $\{a, j\}$, $\{b, c, g, h\}$, and $\{d, e, i, j\}$. However, these learned states do not constitute a control-endogenous latent representation of the MDP. In particular, the resulting transition function is not deterministic (See Figure 2-E). Consequently, the inferred control-endogenous state dynamics are not enough to predict, for example, whether taking the actions sequence L, R, R, R starting in state $\{a, j\}$ will end at state $\{a, j\}$, state $\{b, c, g, h\}$, or state $\{d, e, i, j\}$.

This example demonstrates the non-universality of the multistep-inverse method at learning control-endogenous latent dynamics. In this work, we further develop the theory of endogenous latent dynamics, and demonstrate that combining the multistep-inverse method with a *latent forward dynamics model* is in fact sufficient to learn a control-endogenous latent encoder.

[3]Specifically, the minimum diameter of *any* endogenous representation which meets the assumptions in Appendix B.

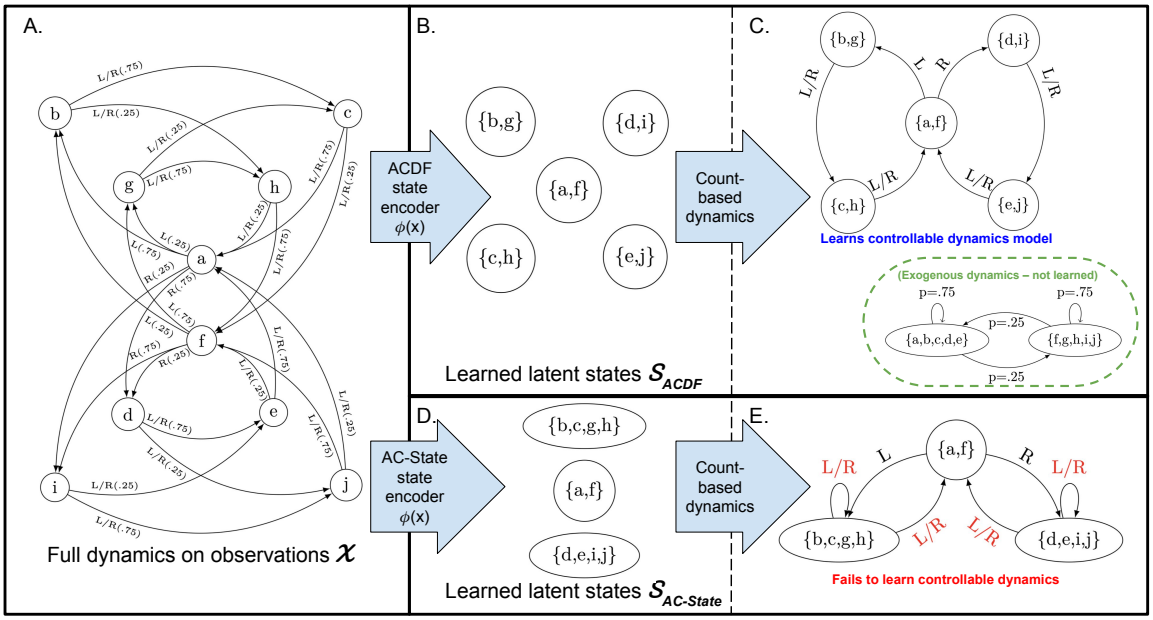


Figure 2: A tabular example where our proposed method ACDF successfully learns a control-endogenous state encoder, while the multistep-inverse method AC-State fails. (A) Full dynamics of the example Ex-BMDP: observed states are $\mathcal{X} = \{a, b..., j\}$ and actions are 'L' and 'R.' Transitions are stochastic: numbers in parentheses after action labels on transitions represent the probability of that transition, conditioned on the action. (B) Encoded latent states $\phi(x) \in \mathcal{S}$, where ϕ is the encoder learned using our proposed method, "ACDF." For example, ϕ maps the observed states b and g to the same latent state in $\mathcal{S}$. (C) Dynamics on the encoded latent states $\mathcal{S}$. The dynamics are deterministic, and capture the full agent-controllable factor of the state. Once ϕ is learned, these dynamics can be inferred from transition data by simple counting. The agent-*independent* exogenous dynamics are shown in the inset: these dynamics are not learned by our method. (D) Encoded latent states produced by the encoder ϕ output by the AC-State algorithm (Lamb et al., 2022). (E) The encoded latent states learned by AC-State are *incorrect*: the encoding conflates states with different forward dynamics, resulting in under-determined transitions between latent states.

3 Guaranteed Learning of Control-Endogenous Dynamics

In this section, we propose a modified loss function to replace $\mathcal{L}_{\text{AC-State}}$, for which we prove (in Appendix E) that any minimum is a correct control-endogenous latent representation. Further, we show that the minimum-range ϕ_θ which minimizes our loss function is a *minimal-state* control-endogenous latent representation. We call our method **ACDF**, or **AC**-State+**D**'+**F**orward. The loss function is given as follows:

$$\begin{aligned}\mathcal{L}_{\text{ACDF}}(\phi_\theta) := &\min_f \mathop{\mathbb{E}}_{k\sim\{1,\ldots,D'\}} \mathop{\mathbb{E}}_{(x_t,a_t,x_{t+k})} -\log(f_{a_t}(\phi_\theta(x_t), \phi_\theta(x_{t+k}); k)) \\ &+ \min_g \mathop{\mathbb{E}}_{(x_t,a_t,x_{t+1})} -\log(g_{\phi_\theta(x_{t+1})}(\phi_\theta(x_t), a_t)).\end{aligned} \tag{3}$$

Where, relative to $\mathcal{L}_{\text{AC-State}}$:

- We have replaced the upper-bound on the control-endogenous diameter, D, with a new quantity D', to be defined below. Note that if we are only given the diameter D, we can use an upper-bound $D' := 2D^2 + D$, which is tight up to a constant multiple in the worst case.

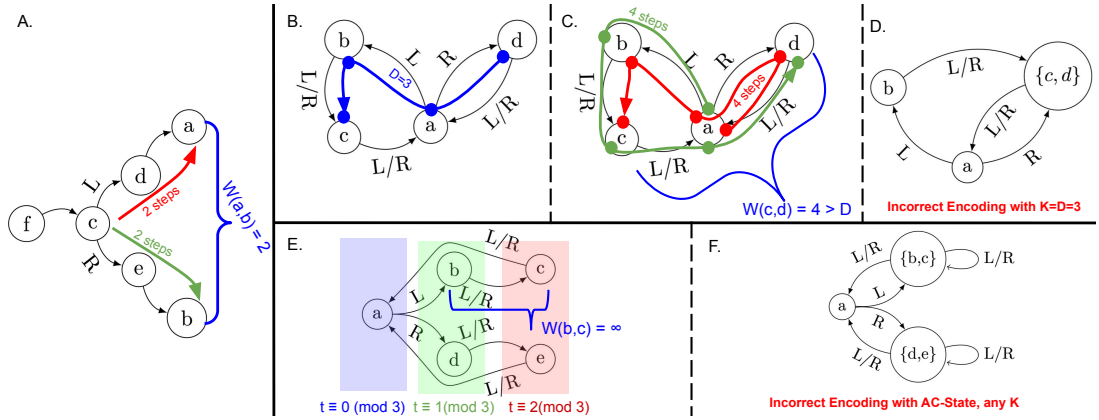


Figure 3: A. Example of the *witness distance* $W(a, b)$. B-D. Witness distance can be greater than D, leading AC-State to fail. E-F. Witness distance can be *infinite* if the dynamics are periodic, which also leads to AC-State failures. (See text of Section 3.1.)

- We have added a *latent forward dynamics model* g over the learned latent states. This model takes the encoded endogenous latent state of an observation x_t, and the action a_t, and returns a normalized probability distribution over the (discrete) encoded latent states. The model is trained to predict the next latent state $\phi_\theta(x_{t+1})$, which should be a deterministic function of the previous state and the chosen action, and is optimized jointly with ϕ_θ.

We explain the logic behind these two modifications below:

3.1 Why D Steps is Insufficient for Multi-Step Dynamics

We assume, as Lamb et al. (2022) does, that the Ex-BMDP admits a latent representation such that the control-endogenous dynamics have diameter upper-bounded by some D, as described in Section 1. In order to introduce our alternative bound D', we first define a few quantities. For a given endogenous representation with states $\mathcal{S}$, we define the *witness distance* $W(a, b)$ between states $a, b \in \mathcal{S}$ as the minimum number of steps k such that there exists some *witness state* $c \in \mathcal{S}$ such that a path of length *exactly* k exists from c to a and also from c to b (See Figure 3-A). Note that this quantity may be infinite even for a bounded-diameter graph, if the dynamics are periodic (in which case $W(a, b) = \infty$, see Figure 3-E). The *witness distance of an endogenous representation* $W(\phi)$ is the *maximum finite* $W(a, b)$, for any pair of states a, b in the endogenous representation. Finally, the quantity D' is defined as any upper bound on the witness distance $W(\phi)$.[4]

The proofs in Lamb et al. (2022) implicitly assume that the witness distance $W(a, b)$ of any pair of states $a, b \in \mathcal{S}$ is upper-bounded by the diameter D of the transition graph. To briefly sketch the main proof in Lamb et al. (2022), let $W(a, b) = k \leq D$ with witness state c. Assume $s_t = c$ and $s_{t+k} \in \{a, b\}$, and we wish to accurately predict a_t given x_t and x_{t+k} (in order to minimize the k-step inverse loss, which is part of our loss function because $k \leq D$, by assumption). Being able to determine from x_{t+k} whether $s_{t+k} = a$ or $s_{t+k} = b$ is *guaranteed to* help us make this prediction, because the sets of possible values that a_t can take will be *disjoint* depending on if $s_{t+k} = a$ or $s_{t+k} = b$. (Otherwise, the witness distance would be most $k - 1$, because a and b could both be reached from s_{t+1}, leading to a contradiction.)

However, it is *not* true in general that $W(a, b) \leq D$. Figure 3-B,C,D gives an example on a simple four-state graph, where $D = 3$, but the largest witness distance $W(c, d) = 4$. The AC-State loss with $K = D = 3$ will learn an encoder that fails to distinguish all of the endogenous latent states.

[4]More precisely, it is defined as any upper bound on the *minimum value* of the witness distance $W(\phi)$ over the set of ϕ's that are minimum-$|S|$, finite-diameter endogenous latent representations of the Ex-BMDP.

To analyse this example: we can't infer the action a_t from s_t and s_{t+k} if s_t is any state other than a, because for all other states, the actions L and R have the same effect. From a, the states reachable in exactly $k = 1$ step are $\{b, d\}$; in $k = 2$ steps are $\{c, a\}$; and in $k = 3$ steps are $\{a, b, d\}$. In particular, when the classifier f is provided with the value of $k \in \{1, 2, 3\}$, there is *no need for the encoder to be able to distinguish* the state c from the state d.

AC-State will therefore produce *incorrect* learned endogenous latent dynamics for an Ex-BMDP with these true endogenous dynamics (Figure 3-D). This learned representation is incorrect because it is in fact *controllable* whether the agent is in state c or state d four steps after being in state a (as seen in 3-C), but the AC-State dynamics do not capture this controllablility.

It also may be *impossible* to reach two states from the same state in exactly the same number of steps, leading to an *infinite* witness distance. This case occurs if the endogenous dynamics are periodic. AC-State may then fail by conflating states belonging to different cyclic classes. (See Figure 3-E,F.)

3.1.1 $D' := 2D^2 + D$ is a Tight Upper-Bound

In Appendix D, we show that if the witness distance $W(a, b)$ between any two states is finite, then it is upper-bounded by $D' := 2D^2 + D$. Therefore, if all pairs of endogenous states have finite witness distance between them, and D (or an upper-bound on D) is known, then the multistep-inverse loss with $K \geq 2D^2 + D$ will be sufficient to distinguish all pairs of endogenous states, and hence learn a correct endogenous state encoder.

Furthermore, we show that this bound is tight up to a constant multiplicative factor. We explicitly construct Ex-BMDPs on which AC-State learns an incorrect state encoder for any $K < D^2/2 + O(D)$, for an infinite sequence of arbitrarily-large values of D. If we use both the multistep-inverse loss and a latent forward-dynamics loss, as in the ACDF algorithm (Equation 3), the particular family of Ex-BMDPs we use to derive this lower-bound no longer minimizes the loss. However, through an alternative construction, we can find Ex-BMDPs where the ACDF loss is minimized by an incorrect encoder if $K < D^2/4 + O(D)$. Therefore the upper bound $D' := 2D^2 + D$ is still tight for ACDF.

3.2 Forward Latent Dynamics for Periodic Transition Functions

The above discussion is applicable only to *finite* witness distances. With bounded-diameter endogenous dynamics, we show in Appendix D that the witness distance between two states can be *infinite*, if and only if the endogenous dynamics are *periodic* (meaning that, for some period $p > 1$, each endogenous state $s \in \mathcal{S}$ can only be visited in time intervals that are multiples of p). Then AC-State may fail regardless of the number of steps K used in the multistep inverse dynamics.

However, we show in Appendix E that augmenting the multistep-inverse loss with a latent forward dynamics loss is sufficient to force the encoder to distinguish between states belonging to different cyclic classes of a periodic endogenous MDP. (In brief, either the states have different latent forward dynamics, or else they can be differentiated entirely by a cyclic *exogenous* state factor, so the Ex-BMDP admits a more-minimal endogenous representation.) Thus, we prove that any ϕ which minimizes the ACDF loss is a correct endogenous latent state encoder.

4 Experiments

Numerical Simulation. First, to capture the *statistical* properties of our proposed method, without concerns about optimization or function approximation, we performed numerical simulations on tabular Ex-BMDPs. In these environments, $|\mathcal{X}|$ is small enough that we can consider *all possible* encoders ϕ, and use count-based estimates for the classifier f. Consequently, the only source of error in minimizing $\mathcal{L}_{\text{AC-State}}$ or $\mathcal{L}_{\text{ACDF}}$ is the *sampling error* caused by limited data collection. Full details of the experiments are included in Appendix G, and results are shown in Figure 4. In general, we see that the forward latent dynamics loss in $\mathcal{L}_{\text{ACDF}}$ not only enabled correct inference in periodic

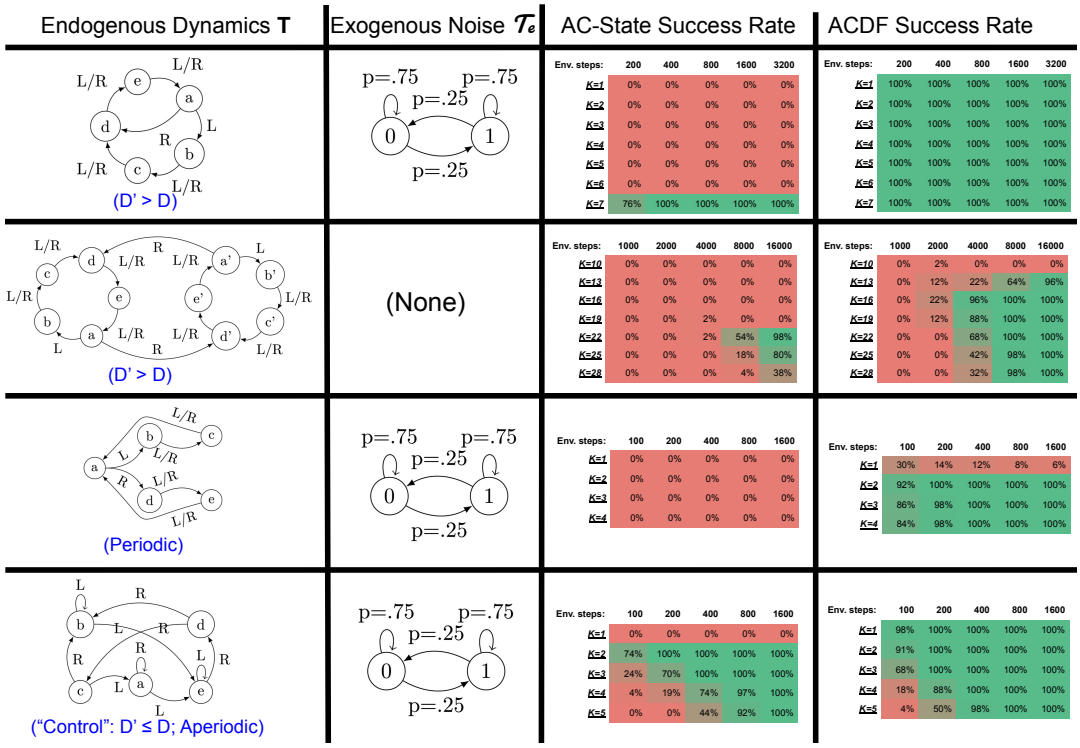


Figure 4: Results of numerical simulation experiments. Four environments are tested, with the dynamics given in the first two columns. For each environment, $|\mathcal{X}| = 10$, and $\mathcal{X}$ is isomorphic to $\mathcal{S} \times \mathcal{E}$. In the last two columns, we show the success rate of each method (AC-State and ACDF) at learning the correct endogenous dynamics over 50 simulations. We show this success rate as a function of the hyperparameter K and the number of environment steps used for learning.

	Baseline/AC-State	Baseline/ACDF	Periodic/AC-State	Periodic/ACDF
Success Rate	20/20 training runs	20/20 " "	1/20 " "	19/20 " "

Table 1: Deep RL Results. Success measured as usability of the final ϕ for open-loop planning.

examples, but also lowered the value of K necessary to learn the dynamics, and additionally yielded improved sample-efficiency even when it wasn't strictly necessary – as in the "control" example.

Deep Reinforcement Learning. We also ran deep-RL experiments on two gridworld-like environments with image observations. The first "baseline" environment, from Lamb et al. (2022)'s released code, consists of nine copies of a four-room maze, with the ego-agent in one maze and random "distractor" agents in the others as exogenous noise. The second environment is constructed similarly, but with dynamics designed to be *periodic*. Brief results for optimized hyperparameters are in Table 1, with more details and results in Appendix H. ACDF preserved performance on the baseline task while also more consistently learning a correct encoder for the periodic task.

5 Related Works

In the area of learning latent representations for reinforcement learning problems, there are numerous ways of defining the "purpose" of the representation – i.e., what information should ideally be included or excluded from the representation. Our work specifically focuses on the Ex-BMDP formulation given in Section 1, which was first proposed by Efroni et al. (2022b), and later studied by Lamb et al. (2022). Note that the setting considered by Efroni et al. (2022b) is *time-inhomogeneous*: the Ex-BMDP is assumed to progress for a finite number of steps H from a (near) deterministic

control-endogenous start state, and the algorithm learns a *different* state representation for *each* time-step. Consequently, number of states needed to represent the dynamics is potentially greatly increased, and this setting does not allow for generalization to long sequences, or encoding when the current time step is not known. In contrast, Lamb et al. (2022) and this work consider the infinite-time-horizon Ex-BMDP. Note that because AC-State is the only prior work to our knowledge that considers the infinite-time-horizon finite-$\mathcal{S}$ Ex-BMDP explicitly, we compare to it as our sole baseline in experiments. (While Wang et al. (2022) considers a similar formulation, with continuous latent states $\mathcal{S}$, the approach proposed in Wang et al. (2022) requires learning a *generative model* of the *complete* system state $\mathcal{X}$, explicitly modeling even the *exogenous* state $\mathcal{E}$ of the system. Kooi et al. (2023) also requires explicitly modeling both the "controllable" and "uncontrollable" latent dynamics, and also uses a somewhat different definition for how these factors relate than in the Ex-BMDP framework.)

Other works present multi-step inverse methods learning for *timestep-dependent* dynamics, including by Mhammedi et al. (2023), which allows for nondeterministic endogenous dynamics, and Efroni et al. (2022a), which considers an explicitly factored state. Multi-step inverse methods have also been used in the continuous-state setting. Mhammedi et al. (2020) considers the special case of linear control-endogenous dynamics, and derives theoretical guarantees in this setting. More recently, Islam et al. (2023) and Koul et al. (2023) have used multi-step inverse methods as an empirical technique for representation learning under continuous latent states. Note that while Koul et al. (2023) does learn a forward dynamics model, the forward-dynamics loss is not used to train the state encoder: it is only used for planning.

Aside from the Ex-BMDP framework and multi-step inverse methods, other methods have been proposed to learn compact relevant state representations. Efroni et al. (2022b) discusses how several of these classes of methods will sometimes include control-exogenous noise into the latent representation, or fail to include control-endogenous information. Techniques such as Deep Bisimulation for Control (Zhang et al., 2020) and DeepMDP (Gelada et al., 2019) ultimately rely on an external reward signal to determine what features are relevant for control, and so may fail to represent controllable aspects of the environment that do not affect the training reward. Misra et al. (2020) demonstrates that one-step inverse dynamics, such as those used empirically by Pathak et al. (2017), are insufficient for learning control-endogenous dynamics. Techniques based on "compressing" states, e.g. through auto-encoders (Hafner et al., 2019) may also include exogenous information.

Hutter & Hansen (2022) is another work that examines the limitations of inverse models. However, it explores the conditions under which an inverse dynamics model on an MDP is sufficient to uniquely learn the transition function. It does not consider endogenous state representations or the "block" setting. Note that the problem of learning an endogenous encoder is somewhat "easier" than learning a forward dynamics model: in the setting we consider, the forward dynamics on the endogenous states are to be learned directly from samples of the transition function *after* learning the encoder. By contrast, Hutter & Hansen (2022) considers the problem of inferring the transition function directly from the inverse dynamics *alone* – the conditions in which these tasks are possible may differ.

6 Limitations and Future Work

One limitation to our work is that the cases where AC-State fails are in some sense "edge cases." In particular, if the endogenous dynamics have even a single state with a transition to itself, then AC-State with $K = D$ should succeed (see Appendix F.2.) However, in many real-world environments, staying in place forever is not possible, so ACDF may still be useful. In addition, we have shown that adding a forward dynamics loss can improve sample efficiency and reduce the dependence on K, even when ACDF is not strictly necessary. A general limitation of the Ex-BMDP model is that it assumes full observation of the state. However, Wu et al. (2023) has extended AC-State to partially-observed environments: this extension can be straightforwardly adapted to ACDF. In Appendix F, we discuss some alternative approaches we initially considered for "fixing" AC-State, which we determine do not work generally; however, this discussion may provide inspiration for future algorithms.

Acknowledgements

A portion of this research has taken place in the Learning Agents Research Group (LARG) at UT Austin. LARG research is supported in part by NSF (FAIN-2019844, NRT-2125858), ONR (N00014-18- 2243), ARO (E2061621), Bosch, Lockheed Martin, and UT Austin's Good Systems grand challenge. Peter Stone serves as the Executive Director of Sony AI America and receives financial compensation for this work. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research. Alexander Levine is supported by the NSF Institute for Foundations of Machine Learning. We thank Alex Lamb for releasing the code for the original AC-State paper (Lamb et al., 2022) at our request.

References

Alfred Brauer. On a problem of partitions. *American Journal of Mathematics*, 64(1):299–312, 1942. ISSN 00029327, 10806377. URL http://www.jstor.org/stable/2371684.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pp. 1665–1674. PMLR, 2019.

Yonathan Efroni, Dylan J Foster, Dipendra Misra, Akshay Krishnamurthy, and John Langford. Sample-efficient reinforcement learning in the presence of exogenous information. In *Conference on Learning Theory*, pp. 5062–5127. PMLR, 2022a.

Yonathan Efroni, Dipendra Misra, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Provably filtering exogenous distractors using multistep inverse dynamics. In *International Conference on Learning Representations*, 2022b. URL https://openreview.net/forum?id=RQLLzMCefQu.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G. Bellemare. Deepmdp: Learning continuous latent space models for representation learning. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2170–2179. PMLR, 2019. URL http://proceedings.mlr.press/v97/gelada19a.html.

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.

Marcus Hutter and Steven Hansen. Uniqueness and complexity of inverse mdp models, 2022.

Riashat Islam, Manan Tomar, Alex Lamb, Yonathan Efroni, Hongyu Zang, Aniket Rajiv Didolkar, Dipendra Misra, Xin Li, Harm Van Seijen, Remi Tachet Des Combes, and John Langford. Principled offline RL in the presence of rich exogenous information. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 14390–14421. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/islam23a.html.

Jacob Eeuwe Kooi, Mark Hoogendoorn, and Vincent Francois-Lavet. Interpretable (un)controllable features in MDP's. In *Sixteenth European Workshop on Reinforcement Learning*, 2023. URL https://openreview.net/forum?id=VZFgkZV3a5.

Anurag Koul, Shivakanth Sujit, Shaoru Chen, Ben Evans, Lili Wu, Byron Xu, Rajan Chari, Riashat Islam, Raihan Seraj, Yonathan Efroni, Lekan Molu, Miro Dudik, John Langford, and Alex Lamb. Pclast: Discovering plannable continuous latent states, 2023.

Alex Lamb, Riashat Islam, Yonathan Efroni, Aniket Rajiv Didolkar, Dipendra Misra, Dylan J Foster, Lekan P Molu, Rajan Chari, Akshay Krishnamurthy, and John Langford. Guaranteed discovery of control-endogenous latent states with multi-step inverse models. *Transactions on Machine Learning Research*, 2022.

Bogdan Mazoure, Remi Tachet des Combes, Thang Long Doan, Philip Bachman, and R Devon Hjelm. Deep reinforcement and infomax learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3686–3698. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/26588e932c7ccfa1df309280702fe1b5-Paper.pdf.

Zakaria Mhammedi, Dylan J Foster, Max Simchowitz, Dipendra Misra, Wen Sun, Akshay Krishnamurthy, Alexander Rakhlin, and John Langford. Learning the linear quadratic regulator from nonlinear observations. *Advances in Neural Information Processing Systems*, 33:14532–14543, 2020.

Zakaria Mhammedi, Dylan J Foster, and Alexander Rakhlin. Representation learning with multi-step inverse kinematics: An efficient and optimal approach to rich-observation rl. *arXiv preprint arXiv:2304.05889*, 2023.

Dipendra Misra, Mikael Henaff, Akshay Krishnamurthy, and John Langford. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pp. 6961–6971. PMLR, 2020.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, pp. 2778–2787. PMLR, 2017.

Peter Perkins. A theorem on regular matrices. *Pacific J. Math.*, 11(4):1529–1533, 1961.

DHJ Polymath. Variants of the selberg sieve, and bounded intervals containing many primes. *Research in the Mathematical sciences*, 1:1–83, 2014.

Tongzhou Wang, Simon Du, Antonio Torralba, Phillip Isola, Amy Zhang, and Yuandong Tian. Denoised mdps: Learning world models better than the world itself. In *International Conference on Machine Learning*, pp. 22591–22612. PMLR, 2022.

Lili Wu, Ben Evans, Riashat Islam, Raihan Seraj, Yonathan Efroni, and Alex Lamb. Agent-centric state discovery for finite-memory POMDPs. In *NeurIPS 2023 Workshop on Generalization in Planning*, 2023. URL https://openreview.net/forum?id=tKTTvWOzsZ.

Amy Zhang, Rowan Thomas McAllister, Roberto Calandra, Yarin Gal, and Sergey Levine. Learning invariant representations for reinforcement learning without reconstruction. In *International Conference on Learning Representations*, 2020.

A A Note on Terminology

Due to assumptions (1), (4), and (5) in the technical assumptions section below (Section B), when discussing MDPs (including Ex-BMDPs) in these appendices, we will most often be referring to finite MDPs under stationary behavior policies that assign nonzero probability to each possible action. As a consequence, properties of the induced Markov chain like periodicity, irreducibly, cyclic classes, etc. will be invariant to the particular choice of behavior policy, and will only depend on the structure of the MDP. Therefore, when unambiguous, we will sometimes refer to, for instance,"the periodicity of the MDP", as shorthand to mean "the periodicity of the induced Markov chain of the MDP under any policy that assigns nonzero probability to each action." (We avoid this shorthand in major theorem statements, such as the statement of Theorem D.1.)

B Technical Assumptions

In this section, we discuss the technical assumptions we make on the Ex-BMDP model, the behavioral policy, and on the collection of data. We also note where these assumptions differ from those of Lamb et al. (2022). Recall that, by the definition of the Ex-BMDP, there must exist *at least one* endogenous state representation, which we call $s^* \in \mathcal{S}^*$, with deterministic dynamics T^* and a corresponding exogenous state representation, which we will call $e^* \in \mathcal{E}^*$. Let the endogenous encoder ϕ^* map $\mathcal{X}$ to $\mathcal{S}^*$, and exogenous encoder ϕ_e^* map $\mathcal{X}$ to $\mathcal{E}^*$. Specifically, we will use these symbols to refer to some correct endogenous state representation which has minimal $|\mathcal{S}|$ among the set of correct endogenous state representations.

While this representation is not necessarily unique (as we show in Appendix C) it will still be useful to refer to it. In particular, some of our assumptions are in terms of this representation. When we refer to an assumption about $\mathcal{S}^*$, $\mathcal{E}^*$, etc., unless otherwise specified, we mean that there exists at least one minimal-endogenous-state decomposition of the Ex-BMDP for which all of these assumptions simultaneously hold.

We first give a brief statement of our assumptions, and then re-state them with more thorough discussion.

Brief Assumptions:

1. In the proofs of the correctness of ACDF, we assume that $\mathcal{X}$ **is finite.**
2. We assume that **for some correct minimal-state endogenous encoder ϕ^*, the endogenous latent dynamics T^* have diameter bounded by D.**
3. We assume that, for the exogenous encoder ϕ_e^* corresponding to the above ϕ^*, **there are no transient exogenous latent states in $\mathcal{E}^*$.**
4. We assume that, for the above-mentioned endogenous encoder ϕ^*, **the behavioral policy $\pi(x)$ used to collect data depends only on $\phi^*(x)$, and ignores exogenous noise $\mathcal{E}^*$.**
5. **Coverage assumptions: For $k \leq D', x, x' \in \mathcal{X}, a \in \mathcal{A}$, if x' is reachable from x, a in exactly k steps, then we assume that we will sample $(x_t = x, a_t = a, x_{t+k} = x')$ with fixed, finite, nonzero probability.** Let $\mathcal{D}_{(k)}$ be the distribution with this property from which $(x_t = x, a_t = a, x_{t+k} = x')$ are sampled when computing the expectation in Equation 3.

We now discuss each of these assumptions in more detail:

1. In the proofs of the correctness of ACDF, we assume that $\mathcal{X}$ **is finite.** This assumption is necessary because otherwise, some observations $x \in \mathcal{X}$ occur with infinitesimal probability, so an incorrect encoding $\phi(x)$ on such observations would not have a finite effect on the

overall loss. (However, in practice, ϕ is a neural network with the capacity to generalize to new observations in a continuous space: this requirement should be thought of as more of a technical limitation of our proofs rather than a real constraint.)

2. Similar to the "bounded diameter" assumption of Lamb et al. (2022), we assume that **for some correct minimal-state endogenous encoder ϕ^*, the endogenous latent dynamics T^* have diameter bounded by D.**

3. We assume that, for the exogenous encoder ϕ_e^* corresponding to the above ϕ^*, **there are no transient endogenous latent states in $\mathcal{E}^*$.** This assumption is necessary because it might not be possible to uniquely determine the endogenous state of an observation if the observation's exogenous state only occurs in the first few steps of a trajectory. (For example, if two endogenous states have the same forward dynamics, it would be impossible to uniquely assign an observation x to one state or the other, if $\phi_e^*(x)$ only occurs at $t = 0$.) Note that as a consequence of this assumption, the states of $\mathcal{E}^*$ can be partitioned into some number of closed, recurrent communicating classes, which each may be periodic or aperiodic. Additionally, because $\mathcal{X}$ is finite, we can assume that $\mathcal{E}^*$ is also finite.

4. We assume that, for the above-mentioned endogenous encoder ϕ^*, **the behavioral policy $\pi(x)$ used to collect data depends only on $\phi^*(x)$, and ignores exogenous noise $\mathcal{E}^*$.** Lamb et al. (2022) also makes this assumption explicitly. While this assumption may seem difficult to meet, because it seemingly requires prior knowledge of ϕ^*, we note that a policy that takes actions in $\mathcal{A}$ according to a fixed distribution at all time-steps (such as a **uniform random policy**) will meet this requirement.

5. **Coverage/Initial Distribution assumptions:** We only provide an asymptotic, rather than statistical, analysis. In other words, our results are only proven to hold for the *population* expectation in Equation 3, which occurs in the limit as samples approach infinity. However, me must still be explicit about the distribution from which we draw tuples (x_t, a_t, x_{t+k}) (and (x_t, a_t, x_{t+1})). To summarize, **for $k \leq D', x, x' \in \mathcal{X}, a \in \mathcal{A}$, if x' is reachable from x, a in exactly k steps, then we assume that we will sample $(x_t = x, a_t = a, x_{t+k} = x')$ with fixed, finite, nonzero probability.** We discuss some implications below:

   - **Unlike Lamb et al. (2022), we do not assume a single trajectory**. Lamb et al. (2022) at one point mentions that data is assumed to be collected in a single, long trajectory, which follows dynamics that have a stationary distribution which assigns finite probability to all *endogenous* states. However, such a sampling process does not necessarily give adequate coverage, because neither we nor Lamb et al. (2022) explicitly assume that the *overall* dynamics $\mathcal{T}$ on the observations $\mathcal{X}$ are irreducible. In particular, $\mathcal{T}$ may be reducible if either the exogenous dynamics $\mathcal{T}_e^*$ are reducible, or if both $\mathcal{T}_e^*$ and the endogenous dynamics are periodic with the same period (a case we discuss further below). **We therefore assume at least one trajectory for each communicating class in $\mathcal{X}$.** Because $\mathcal{X}$ is assumed finite, we can assume a finite number of these communicating classes. When we consider the "infinite sample" limit, we can assume that the length of *each* of these trajectories goes to infinity (and therefore the time-averaged state visitation approaches a stationary distribution).
   - Because (under any policy that assigns nonzero probability to every action) the endogenous dynamics are irreducible while the exogenous dynamics only have recurrent communicating classes, it follows that each communicating class of $\mathcal{X}$ is finite, closed and recurrent. Therefore **any stationary policy that assigns nonzero probability to each action while in each endogenous state $s^* \in \mathcal{S}^*$ will, averaging over time, approach a stationary distribution assigning finite, nonzero probability to each x in the communicating class in $\mathcal{X}$ that the chain started in.** Because chains will reach each x with finite, nonzero probability and the execute each sequence of actions of finite length with finite probability, they **will therefore sample**

each possible $(x_t = x, a_t = a, x_{t+k} = x')$ with fixed finite probability. Any collection of such chains is a therefore valid data-collection mechanism.

- **Unlike Lamb et al. (2022), we do not assume that the initial exogenous and endogenous states are independent.** Lamb et al. (2022) assumes that the initial distribution of the Ex-BMDP $\mathcal{D}_0$ is such that $\phi^*(x_0)$ and $\phi_e^*(x_0)$ are distributed independently. This assumption seems overly-strict, given that, as discussed above, Lamb et al. (2022) *also* assumes that collecting a single trajectory is sufficient. (If a single episode is sufficient to learn the dynamics, how can the *distribution* of x_0 possibly matter?) We do *not* make this independence assumption, rather, **we define $\mathcal{D}_0$ to be a distribution over $\mathcal{X}$ directly.**[5] A particular consequence of *not* making this independence assumption is that **we do not assume that all pairs (s^*, e^*) necessarily correspond to an observation in $\mathcal{X}$.** In other words, **we assume coverage over $\mathcal{X}$, not over $\mathcal{S}^* \times \mathcal{E}^*$.** This consequence comes into play when both the endogenous and exogenous dynamics are periodic, with the same period. We discuss the implications in Section B.1 below.
- **A note about periodicity in Lamb et al. (2022):** One of the major claims of this paper is that Lamb et al. (2022) incorrectly handles Ex-BMDPs with periodic endogenous dynamics. Note that Lamb et al. (2022) makes the explicit assumption that "Markov chain $\mathcal{T}_D$ has a stationary distribution μ_D such that $\mu_D(s, a) > 0$ and $\pi_D(a|s) \geq \pi_{\min}$, for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$", where "$\mathcal{T}_D(s'|s)$ [is] the Markov chain induced on the control-endogenous state space by executing the policy π_D by which AC-State collects the data." We wish to emphasise that this assumption *should not* preclude Ex-BMDPs with periodic dynamics. In particular, note that periodic Markov chains (like $\mathcal{T}_D$ would be in the case of periodic dynamics) can indeed have stationary distributions that assign nonzero probabilities to all states. Recall that a stationary distribution is merely any distribution μ for which $\mu P = \mu$, where P is the transition matrix. The existence of such a distribution is a property only of the transition function P, not the initial distribution. Therefore, even if a periodic Markov chain has a fixed initial state x_0, and therefore the distribution of the random variable x_t (for any fixed t) will only have nonzero probability on states of one cyclic class, it is *still the case* that the Markov chain can have a *stationary distribution* that assigns probability to every state. Therefore Lamb et al. (2022) is not "off the hook" for periodic Ex-BMDPs. (Moreover, the periodicity or aperiodicity is never discussed explicitly in Lamb et al. (2022), and the lines quoted above are part of a discussion about the behavioral policy π_D, *not* about the structure of the Ex-BMDP. Therefore it is doubtful this assumption was intended in any way to exclude periodic dynamics.)

B.1 Implications of not assuming independence in initial exogenous and endogenous states

Above, we mentioned that, unlike Lamb et al. (2022), we do not assume that the initial exogenous and endogenous states are distributed independently. This fact matters when both the endogenous and exogenous dynamics are periodic, with the same period. Assume that this period is k, and the exogenous dynamics are irreducible, such that $\mathcal{S}^*$ has cyclic classes $\{\mathcal{S}_0^*, ..., \mathcal{S}_{k-1}^*\}$ and $\mathcal{E}^*$ has cyclic classes $\{\mathcal{E}_0^*, ..., \mathcal{E}_{k-1}^*\}$. If the support of $\mathcal{D}_0$ only includes states x_0 such that, for example $(\phi^*(x_0), \phi_e^*(x_0))) \in \mathcal{S}_0^* \times \mathcal{E}_1^*$, then the only states which will *ever* be reachable will be states x such that $(\phi^*(x), \phi_e^*(x)) \in \mathcal{S}_0^* \times \mathcal{E}_1^* \cup \mathcal{S}_1^* \times \mathcal{E}_2^* \cup ... \cup \mathcal{S}_{k-1}^* \times \mathcal{E}_0^*$. Because $\mathcal{X}$ is defined as the set of states that the Ex-BMDP can be in, it is therefore the case that $\forall x \in \mathcal{X}$, $(\phi^*(x), \phi_e^*(x)) \in \mathcal{S}_0^* \times \mathcal{E}_1^* \cup \mathcal{S}_1^* \times \mathcal{E}_2^* \cup ... \cup \mathcal{S}_{k-1}^* \times \mathcal{E}_0^*$. Note that in this case, $\mathcal{X}$ is *not* equivalent to the union of the supports of $\mathcal{Q}^*(s^*, e^*)$, for all $s^* \in \mathcal{S}^*$ and all $e^* \in \mathcal{E}^*$: in fact, in this case, $\mathcal{Q}^*(s^*, e^*)$ is not even well-defined for cases where s^* and e^* cannot co-occur.

[5] We do, however, assume that x_0 is in fact in the support of $Q^*(\phi^*(x_0), \phi_e^*(x_0))$; in other words, that any observation x in the support of $\mathcal{D}_0$ is also in the support of $Q^*(s^*, e^*)$ for *some* choice of (s^*, e^*).

(Also, note that the *initial distribution of the Ex-BMDP* $\mathcal{D}_0$ is a distinct concept from the *initial states used in the sampling trajectories.* The former is a property of the Ex-BMDP itself; the latter is a property of our algorithm.)

This difference in assumptions is important because there exist some Ex-BMDPs where a decomposition with *independent* (s_0, e_0) *exists*, but a more-minimal endogenous representation is possible if we allow for arbitrary initial distributions over $\mathcal{X}$. The ACDF algorithm will find these more-minimal representations. Explicitly, consider the following deterministic Ex-BMDP with $|\mathcal{X}| = 10$ and $\mathcal{A} = \{L, R\}$:

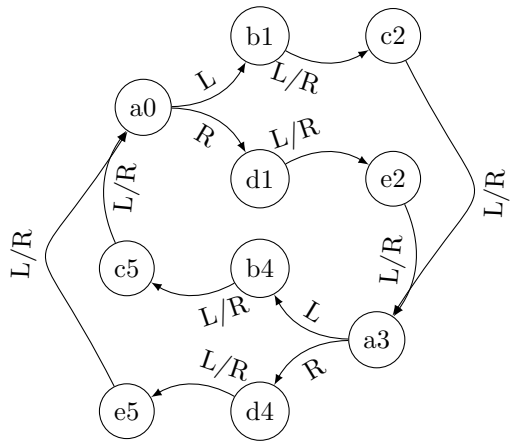
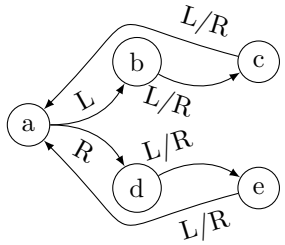


Figure 5: Full Ex-BMDP model of example in Section B.1

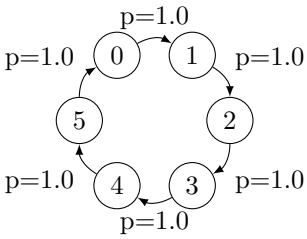
.

For this Ex-BMDP, let the initial state distribution $\mathcal{D}_0$ take value $x_0 = b1$ with probability 0.5, and $x_0 = e2$ with probability 0.5.

The ACDF algorithm will learn the minimal-state endogenous representation of the Ex-BMDP shown in Figure 6, with $|\mathcal{S}| = 5$.



(a) Minimal endogenous dynamics learned by ACDF.



(b) Corresponding exogenous dynamics.

Figure 6: Decomposition of the Ex-BMDP in Figure 5 that would be learned by the ACDF algorithm. As usual, the ACDF algorithm learns an encoder for the endogenous state alone (left); the exogenous dynamics are implicit. (The observation x emitted by latent states (s, e) is given by concatenating the labels of s and e; for example state $x =$ 'c5' is reached when $s =$ 'c' and $e =$ '5'.)

This is a minimal endogenous latent representation, and one can confirm that the endogenous and exogenous latent dynamics together are equivalent to the dynamics shown in Figure 5 on any trajectory that starts on $\mathcal{D}_0$. However, under this decomposition, the initial state distribution $\mathcal{D}_0$ corresponds to a non-independent joint distribution on $\mathcal{S}$ and $\mathcal{E}$: it assigns probability 0.5 to the pair $(s_0 = \text{'}b\text{'}, e_0 = \text{'}1\text{'})$ and probability 0.5 to the pair $(s_0 = \text{'}e\text{'}, e_0 = \text{'}2\text{'})$.

By contrast, consider the *trivial* encoder $\phi'(x) = x$, where $\mathcal{S}' = \mathcal{X}$ and $\mathcal{E}'$ consists of a single state with a self-edge. This is also a valid control-endogenous representation (the transitions on $\mathcal{S}'$ are deterministic, and the (single-state) exogenous dynamics do not depend on actions.) Moreover, s'_0 and e'_0 are independent in $\mathcal{D}_0$ under this decomposition. Therefore, if we were to accept independent initial-state distribution as a *defining feature* of the Ex-BMDP framework, we would be forced to return this trivial encoder with $|\mathcal{X}| = 10$, as the only valid Ex-BMDP decomposition.

We reject this assumption, because we believe that the more-minimal encoder returned by ACDF better captures an intuitive notion of *controllability*: an agent has no control over whether they are in state 'a0' or 'a3' for instance. Further, while it might seem like the fact that $|\mathcal{S} \times \mathcal{E}| > |\mathcal{X}|$ means that our returned encoding is *redundant*, recall that *only the encoding of $\mathcal{S}$ is actually learned.* In fact, loosening the definition of the Ex-BMDP model in this way can only lead to more concise learned representations (because it can only increase the set of valid decompositions of the Ex-BMDP).

Lastly, we note that, while Lamb et al. (2022) may *state* an assumption of decoupled initial exogenous and endogenous states, AC-State also does not enforce in any way that the returned representation will have independent exogenous and endogenous states, even if such a representation is possible. (In fact, on this particular example, AC-State will return an *incorrect* encoding consisting of 3 states with nondeterministic dynamics).

B.1.1 Implications for the structure of $\mathcal{X}$

As discussed above, our lack of an independence assumption on s_0, e_0 means that $\mathcal{X}$ may not contain elements corresponding to all pairs (s, e), for $s \in \mathcal{S}$, $e \in \mathcal{E}$. However, we prove the following lemma which will be useful in showing when such an (s, e) pair does correspond to some element in $\mathcal{X}$.

Lemma B.1. *Consider any policy on $\mathcal{S}^*$ that assigns nonzero probability to all actions (i.e., any valid behavioral policy). Let $s, s' \in \mathcal{S}^*$ and $e, e' \in \mathcal{E}^*$. If (s', e') is reachable from (s, e), then (s, e) is reachable from (s', e'). Consequentially, if (s', e') is reachable from (s, e), and (s', e') corresponds to an observation in $\mathcal{X}$, then (s, e) also corresponds to an observation in $\mathcal{X}$.*

Proof. Note that the dynamics on $\mathcal{S}^*$ are irreducible, and there are no transient states in $\mathcal{E}^*$. We know that e and e' must belong to the same communicating class in $\mathcal{E}^*$, so we can treat the dynamics on $\mathcal{E}^*$ as irreducible without loss of generality.

Let M_s and M_e be the two Markov chain transition matrices. Let k_s be the periodicity of the endogenous dynamics, and k_e be the periodicity of the exogenous dynamics. Note that $M_s^{k_s}$ is ergodic when restricted to the domain of each cyclic class of $\mathcal{S}^*$, and $M_e^{k_e}$ is ergodic when restricted to the domain of each cyclic class of $\mathcal{E}^*$. Then for some n, m, for any $n' \geq n$ and $m' \geq m$, $M_s^{k_s \cdot n'}$ has positive probability between any two states in the same cyclic class in $\mathcal{S}^*$ and $M_e^{k_e \cdot m'}$ has positive probability between any two states in the same cyclic class in $\mathcal{E}^*$. Then in particular, $M_s^{k_s \cdot k_e \cdot m \cdot n}$ has positive probability between any two states in the same cyclic class in $\mathcal{S}^*$, and $M_e^{k_s \cdot k_e \cdot m \cdot n}$ has positive probability between any two states in the same cyclic class in $\mathcal{E}^*$.

Now, let the cyclic classes of e and s be defined as 0 on their respective Markov chains. Let $l((s, e), (s', e'))$ be the length of a path from (s, e) to (s', e'). Now, consider any state (s'', e'') reached by taking $k_e \cdot k_s - (l((s, e), (s', e')) \,\%\, (k_e \cdot k_s))$ steps starting at (s', e'). Then (s'', e'') can be reached in a path of length $l((s, e), (s'', e''))$ from (s, e), where

$$l((s, e), (s'', e'')) = l((s, e), (s', e') + k_e \cdot k_s - (l((s, e), (s', e')) \,\%\, (k_e \cdot k_s)) \equiv 0 \pmod{k_e \cdot k_s} \quad (4)$$

Then s'' belongs to the same cyclic class (0) as s on M_s, and e'' belongs to the same cyclic class (0) as e on M_e. Then $M_s^{k_s \cdot k_e \cdot m \cdot n}$ has positive probability to transition from s'' to s, and $M_e^{k_s \cdot k_e \cdot m \cdot n}$ has positive probability to transition from e'' to e. Then we can reach (s, e) from (s'', e'') in a finite number $(k_s \cdot k_e \cdot m \cdot n)$ of steps. Because we can reach (s'', e'') from (s', e'), this implies that we can reach (s, e) from (s', e'), as desired. As a consequence, if (s', e') corresponds to an observation in $\mathcal{X}$, then the latent dynamics of the Ex-BMDP will eventually reach (s, e) with positive probability, so (s, e) also corresponds to an observation in $\mathcal{X}$. □

C Endogenous Latent Dynamics are not Unique

In this section, we provide some clarifications on the theory of endogenous latent dynamics, which allow us to more clearly state our theoretical results. In their theoretical presentation, Lamb et al. (2022) implicitly assumes that a Ex-BMDP has a *unique* endogenous latent representation; i.e., that a single Ex-BMDP only admits a single (minimal $|\mathcal{S}|$) decomposition of the observation x into endogenous state s and exogenous state e.[6] Consequentially, their theoretical claims are often made in terms of the "ground truth" endogenous latent state. Here, we show that this assumption is unwarranted.

C.1 An Ex-BMDP with Multiple Control-Endogenous Representations

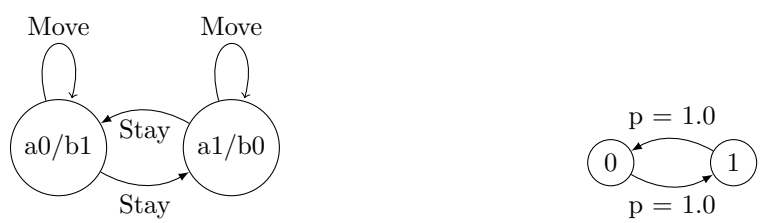
Consider the Ex-BMDP defined by control-endogenous states $\mathcal{S} = \{a, b\}$, exogenous states $\mathcal{E} = \{0, 1\}$, actions $\mathcal{A} = \{\text{Stay}, \text{Move}\}$ and transitions as follows:

(a) Endogenous transitions $T(s, a)$ (b) Exogenous transitions $\mathcal{T}_e(e|e_t)$

Figure 7: An Ex-BMDP with $|\mathcal{S}| = 2$

The observations $x \in \mathcal{X}$ are defined by simply concatenating the endogenous and exogenous state labels: $\mathcal{X} = \{a0, a1, b0, b1\}$.

This Ex-BMDP in fact admits a different control-endogenous state representation, with the same (minimal) endogenous state space size $|\mathcal{S}| = 2$. In particular, consider the representation:



(a) Endogenous transitions $T(s, a)$ (b) Exogenous transitions $\mathcal{T}_e(e|e_t)$

Figure 8: A different decomposition of the same Ex-BMDP, also with $|\mathcal{S}| = 2$

In this decomposition, the observation x is defined such that, for example, if $s = a0/b1$ and $e = 0$, then $x = a0$.

Note that this alternative encoding is not a mere relabeling of the same endogenous states: the pair of observations $\{a0, a1\}$ belong to *different* endogenous states under this representation, although they have the same endogenous state under the first representation. We can confirm that the full MDP (that is, the observed MDP on $\mathcal{X}$) is in fact the same for both:

[6] For example, Lamb et al. (2022) makes claims such as "We present an asymptotic analysis of AC-State showing it recovers the control-endogenous latent state encoder f^*."

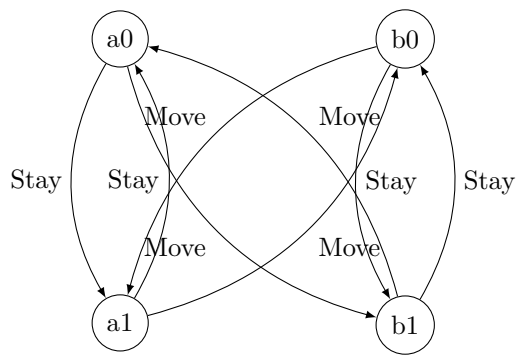


Figure 9: Observed transitions on the full MDP

D Bounds on D' and K

In this section, we prove an upper bound on the maximum witness distance D' in terms of the endogenous transition diameter D. We also show lower bounds on the number of steps K of multi-step inverse needed to correctly learn a dynamics model in terms of D.

Theorem D.1. *Consider a MDP on states $\mathcal{S}$ with deterministic transition function T, and finite diameter D as defined by Assumption 3.1 from Lamb et al. (2022) (reproduced in Section 2 above). Recall the definition of witness distance $W(a, b)$ from Section 3.1. For any $a, b \in \mathcal{S}$, either $W(a, b) \leq 2D^2 + D$, or $W(a, b) = \infty$. Furthermore, $W(a, b) = \infty$ if and only if all Markov chains induced by any policy on T (that assigns nonzero probability to each action) are periodic, and a and b belong to different cyclic classes of such chains.*

Proof. Consider a Markov chain induced by any policy on the transition function T that assigns nonzero probability to each action in each state. Because of the finite diameter, we know that the Markov chain is irreducible. We will initially consider the case that the Markov chain is aperiodic, and later reduce the general case to the aperiodic case.

Our proof technique is inspired by a related result from Perkins (1961), which demonstrated that, for an irreducible aperiodic Markov chain consisting of N states, a path of length exactly $N^2 - 2N + 2$ exists between any pair of states. Building on this technique, our proof additionally takes advantage of the bounded-diameter assumption present in our setting, instead of simply using $|\mathcal{S}|$. Perkins (1961) uses the following theorem sourced from Brauer (1942), which is ultimately attributed to Issai Schur (presented here in the form used in Perkins (1961)):

Theorem (Theorem of Schur)**.** *Consider any set of positive integers $\mathcal{B} = \{b_1, ..., b_k\}$ such that $gcd(\mathcal{B}) = 1$, where $k = |\mathcal{B}|$, b_1 is the smallest integer, and b_k the largest. Then any integer $n \geq (b_1 - 1)(b_k - 1)$ can be represented as a sum in the form $n = \sum_{i=1}^{k} a_i b_i$, where $a_1, ..., a_k$ are non-negative integers.*

For any state $a \in \mathcal{S}$, we know from the definition of aperiodicity that the greatest common denominator of the lengths of all possible self-loops from a to a must be 1. Let $\mathcal{L}_a$ be this complete set of self-loops of a. We will use $|\cdot|$ to denote length of a path (such as a loop). We will show first that there exists a subset $\mathcal{Q}_a \subseteq \mathcal{L}_a$, such that the length of each loop in $\mathcal{Q}_a$ is at most $2D + 1$, the shortest loop in $\mathcal{Q}_a$ has length at most $D + 1$, and such that $\gcd(\{|p| \big| p \in \mathcal{Q}_a\}) = 1$.

Consider the subset $\mathcal{P}_a \subseteq \mathcal{L}_a$ consisting of paths in the form $a \to ... \to a' \to ... \to a$, $\forall a' \in \mathcal{S} \setminus \{a\}$, where the segments $a \to ... \to a'$ take a shortest-possible path from a to a', and conversely, the segments $a' \to ... \to a$ take a shortest-possible path from a' to a. In other words, $\mathcal{P}_a$ is the set of shortest self-loops of a containing each possible other state a'. Let $d(a, a')$ be the length of the shortest path from a to a'. By the bounded-diameter assumption, we know that $|p| = d(a, a') + d(a', a) \leq 2D, \forall p \in \mathcal{P}_a$. Furthermore, there must exist some a' with an edge directly

incident on a: for this a', we have $d(a', a) = 1$, and thus $|p| = d(a, a') + d(a', a) \leq D + 1$. Therefore $\forall p \in \mathcal{P}_a,\ |p| \leq 2D$, and $\exists p \in \mathcal{P}_a,\ |p| \leq D + 1$.

Now, there are two possibilities:

1. $\gcd(\{|p| \big| p \in \mathcal{P}_a\}) = 1$. In this case we tan take $\mathcal{Q}_a = \mathcal{P}_a$

2. $\gcd(\{|p| \big| p \in \mathcal{P}_a\}) = n > 1$. In this case, let m be any factor of n greater than 1. Because $\gcd(\{|p| \big| p \in \mathcal{L}_a\}) = 1$, there must exist some $p' \in \mathcal{L}_a$ such that $m \nmid |p'|$ (m does not divide $|p'|$). (Furthermore, we can assume that the state a does not occur in the middle of p': if we are given such a "double-loop" with $m \nmid |p'|$, then m must not divide the length of either the first segment or the second segment, or both.) Let p'_i be the ith element in the path p' (zero-indexed). Then $p'_0 = p'_{|p'|} = a$. Note that

   - $d(a, p'_{|p'|}) - d(a, p'_0) = 0 - 0 = 0 \equiv 0 \pmod{m}$
   - $|p'| - 0 = |p'| \not\equiv 0 \pmod{m}$

   Therefore, it *cannot* be the case that for all $i \in [0, |p'|)$, $\quad d(a, p'_{i+1}) - d(a, p'_i) \equiv 1 \pmod{m}$. (To confirm this fact, sum both sides of this equivalence over i from 0 to $|p'|-1$.) Fix i so that $p'_{|i|}$ is some state on p' where this equivalence does not hold. (That is, $d(a, p'_{i+1}) - d(a, p'_i) \not\equiv 1 \pmod{m}$.) Now, consider a self-loop of a constructed from the following segments:

   - A shortest-length path from a to p'_i
   - The edge from p'_i to p'_{i+1}
   - A shortest-length path from p'_{i+1} to a.

   Let this loop be known as p'', and note that $|p''| \leq 2D+1$, because it consists of two shortest paths and a single edge. Also, observe that:

$$
\begin{aligned}
d(a, p'_{i+1}) - d(a, p'_i) &\not\equiv 1 \pmod{m} && \text{(As derived above)} && (5)\\
d(a, p'_{i+1}) + d(p'_{i+1}, a) &\equiv 0 \pmod{m} && \text{(A self-loop in } \mathcal{P}_a\text{)} && (6)\\
-d(p'_{i+1}, a) - d(a, p'_i) &\not\equiv 1 \pmod{m} && \text{(Subtract Eq. 6 from Eq. 5)} && (7)\\
d(a, p'_i) + d(p'_{i+1}, a) &\not\equiv -1 \pmod{m} && \text{(Rearrange and negate Eq. 7.)} && (8)\\
d(a, p'_i) + 1 + d(p'_{i+1}, a) &\not\equiv 0 \pmod{m} && \text{(Add 1 to Eq. 8.)} && (9)\\
|p''| &\not\equiv 0 \pmod{m} && \text{(From the definition of } p'' \text{ and Eq. 9.)} && (10)
\end{aligned}
$$

   Note that we can construct such a p'' for each factor m of n. Then we can let $\mathcal{Q}_a$ consist of each loop in $\mathcal{P}_a$, and additionally one of these p'' loops constructed for for each factor m. Then by construction, $\gcd(\{|p| \big| p \in \mathcal{Q}_a\}) = 1$ and $\forall p \in \mathcal{Q}_a,\ |p| \leq 2D + 1$, as desired. (Additionally, $\mathcal{Q}_a$ still contains the loop of length $\leq D + 1$ from $\mathcal{P}_a$, as desired.)

Now, we have constructed $\mathcal{Q}_a$ such that the longest loop in $\mathcal{Q}_a$ has length at most $2D + 1$, while the shortest loop has length at most $D + 1$, and furthermore that the lengths of the loops are relatively prime. Then the Theorem of Schur given above tells us that by taking some combination of self-loops in $\mathcal{Q}_a$ from a in sequence, we can construct a path from a to a of any arbitrary length n, for any $n \geq (2D + 1 - 1)(D + 1 - 1) = 2D^2$. Then for any $b \in \mathcal{S}$, we can construct a path from a to b of length exactly $2D^2 + d(a, b)$ (by self-looping for exactly $2D^2$ steps at a, and then taking the shortest path to b), and also a path from a to a of length exactly $2D^2 + d(a, b)$ (because $2D^2 + d(a, b) \geq 2D^2$, so such a self-loop must exist). Then a can act as a witness for a and b, and $W(a, b) \leq 2D^2 + d(a, b) \leq 2D^2 + D$, as desired.

Finally, we return to the periodic case. Let k be the periodicity of the Markov chain, such that the chain is partitioned into the k cyclic classes $\mathcal{S}_0, ..., \mathcal{S}_{k-1}$, where a state in class $i \pmod{k}$ is always succeeded by a state in class $i + 1 \pmod{k}$. There are two possibilities:

- States a and b belong to the same cyclic class, which we can call $\mathcal{S}_{i_a}$. Then consider the Markov chain raised to the k'th power; that is, the Markov chain induced by taking k steps of the original Markov chain at each step; call this new chain M^k. We will show that M^k is irreducible, bounded-diameter, and aperiodic, and thus we can apply the above aperiodic case to M^k. First, consider any path starting in $\mathcal{S}_{i_a}$ in the original chain, of length $|p|$. Either $k \mid |p|$, in which case it ends in $\mathcal{S}_{i_a}$ and has an equivalent path of length $|p|/k$ in M^k, or $k \nmid |p|$, in which case it does not end in $\mathcal{S}_{i_a}$ and also has no equivalent path in M^k. Thus we see that, if we start in $\mathcal{S}_{i_a}$, M^k forms an irreducible Markov chain on the states of $\mathcal{S}_{i_a}$ alone. Furthermore, M^k on the states $\mathcal{S}_{i_a}$ has diameter at most $\lfloor D/k \rfloor$ (because there is a path of length of most D between every pair of states in $\mathcal{S}_{i_a}$ in the original Markov chain). Finally, M^k on the states $\mathcal{S}_{i_a}$ is aperiodic: to confirm this fact, note that by the definition of periodicity, we have that in the original Markov chain, the g.c.d. of the lengths of all self-loops from any particular state in $\mathcal{S}_{i_a}$ is k. All of these self-loops will still exist in M^k, but their lengths will be divided by k. Therefore, the g.c.d. of the lengths of the self-loops will be 1. This fact directly implies that M^k on $\mathcal{S}_{i_a}$ is aperiodic. Now, because M^k on $\mathcal{S}_{i_a}$ is irreducible and aperiodic with bounded diameter $\lfloor D/k \rfloor$, we can apply the aperiodic case above to M^k. Then, in M^k on $\mathcal{S}_{i_a}$, we have that $W(a,b) \leq 2\lfloor D/k \rfloor^2 + \lfloor D/k \rfloor \leq 2D^2/k^2 + D/k$. The paths from the witness state to a and b will also exist in the original MDP, but will be longer by a factor of k. Therefore the witness distance in the original MDP will be $W(a,b) \leq 2D^2/k + D \leq 2D^2 + D$.

- States a and b belong to different cyclic classes, which we can call $\mathcal{S}_{i_a}$ and $\mathcal{S}_{i_b}$, where $i_a \not\equiv i_b \pmod k$. In this case, for any witness state c in any $\mathcal{S}_{i_c}$, note that any path from c to a will have length congruent to $i_a - i_c \pmod k$, while any path from c to b will have length congruent to $i_b - i_c \pmod k$. Because $i_a \not\equiv i_b \pmod k$, we have that $i_a - i_c \not\equiv i_b - i_c \pmod k$, so no path from c to a can be the same length as any path from c to b. Then $W(a,b) = \infty$. Note that this case is the **only** case in which $W(a,b) = \infty$.

□

The above result shows that the maximum witness distance D' is upper-bounded by $2D^2 + D$, and thus, if we are given diameter D, we can safely set the hyperparameter $K = 2D^2 + D$ and be guaranteed that ACDF will correctly discover a endogenous state representation.

We now show two kinds of lower bounds on K, the number of multi-step inverse steps *actually* needed to learn the endogenous state:

- A bound for the AC-State loss. For this bound, we assume that the dynamics are aperiodic; thus $K = 2D^2 + D$ steps are in fact sufficient to find a correct endogenous state representation.
- A bound for the ACDF loss.

These results are proven by construction: we will explicitly construct examples of dynamics for which AC-State/ACDF fail when K is too small.

We first give the following useful lemma:

Lemma D.2. *Consider an Ex-BMDP defined on states indexed as $\mathcal{X} = \{0, ..., |\mathcal{X}| - 1\}$, such that the dynamics of the Ex-BMDP are deterministic, and such that any Markov chain induced by any policy on the Ex-BMDP (that puts nonzero probability on every action) is irreducible and aperiodic. Then the only valid endogenous latent encoder (up to a relabeling permutations) is the trivial one: $\phi(x) = x$. That is to say, the endogenous state s is the full state x, and the endogenous transition function is the same as the full Ex-BMDP transition function.*

Proof. We proceed as follows:

- Consider any pair of endogenous state s and exogenous state e. By the block assumption, each observed state x corresponds to exactly one such pair, and any two observed states x, x' which correspond to the same pair (s, e) must have in-edges in M from exactly the same set of states in $\mathcal{X}$, with exactly the same actions for each edge, with a fixed probability ratio. Because the Ex-BMDP is deterministic, there are no such pairs, and it immediately follows that each pair (s, e) corresponds to exactly one x, and vice-versa.
- Because the full dynamics of $x = (s, e)$ are deterministic, and because the dynamics of s are deterministic (by definition), it immediately follows that the dynamics of e must be deterministic.
- Given that the dynamics of e are deterministic and by definition do not depend on actions, $\mathcal{E}$ and $\mathcal{T}_e$ can only consist of some set of disjoint cycles of states. However, if there are multiple such cycles, then under any choice of policy, any Markov chain on $x = (s, e)$ would have multiple communicating classes (because states $x = (s, e)$ with values of e on different cycles would be inaccessible from each other), and would therefore not be irreducible. Therefore $\mathcal{E}$ can only consist of a single cycle. However, this cycle can in fact only consist of a single state: note that the length of any self-loop of any state $x = (s, e)$ will by divisible by the cycle length of $\mathcal{E} = |\mathcal{E}|$. Because the Markov chain is aperiodic, the g.c.d of such loops must be 1. Then $|\mathcal{E}|$ must be equal to 1.

Therefore, each x corresponds one-to-one with a endogenous-exogenous pair (s, e), and there is only a single exogenous state e. So each x corresponds one-to-one with an endogeous state s, as desired. □

We now present our lower-bound results:

Proposition D.3. *$\forall D$, there exists an Ex-BMDP such that a minimal endogenous latent dynamics has a diameter of D and is aperiodic, and such that the AC-State algorithm given by Equation 2 with a uniform exploration policy will return an encoder that does **not** produce a valid endogenous state representation using $K \leq h(D)$ -step inverse dynamics, where $h(D) \in \Omega(D^2)$. Here, we are using the Hardy-Littlewood Ω notation (not to be confused with Knuth's Ω), and specifically $h(D) \sim D^2/2$, in the sense that:*

$$\limsup_{D \to \infty} \frac{h(D)}{\frac{1}{2}D^2} = 1 \tag{11}$$

Proof. Consider two arbitrary primes p, q, with $p < q$, and let these primes define a deterministic Ex-BMDP M with transition function $T(x, a)$ as follows:

- q states, labeled $\mathcal{X} = \{0, ..., q-1\}$
- Two actions, $\{L, R\}$
- $T(s, a)$ defined as:
  - $T(0, L) = 1$
  - $T(0, R) = q - p + 1$
  - $\forall x \in \mathcal{X}, x \neq 0$: $T(x, L) = T(x, R) = (x + 1) \,\%\, q$, where % is the modulo operator.

We first note that for this Ex-BMDP, under any policy that assigns nonzero probability to all actions, the resulting Markov chain is irreducible and aperiodic. For irreducibility, simply note that any state can be reached from any other state (i.e., any state b can be reached from any state a by simply taking action L $(b - a) \,\%\, q$ times). For aperiodicity, note that the state $0 \in \mathcal{X}$ has a self-loop of length q (by taking L and then any other $q - 1$ actions) and a self-loop of length p (by taking R and then any other $p - 1$ actions). Because these are both prime, the g.c.d. of the lengths of all of state 0's self-loops is 1, so the Markov chain is aperiodic. Then, by applying Lemma D.2, we know

that the *only* valid endogenous latent encoder (up to a relabeling permutations) is the trivial one: $\phi(x) = x$. We can thus regard the states $\mathcal{S} = \mathcal{X}$ as the only "true" endogenous states: any encoder which maps any two states in $\mathcal{X}$ to the same endogenous state must be incorrect.

Now, consider the AC-State prediction task of determining a_t from $\phi(x_t)$, $\phi(x_{t+k})$, and k. We will assume that ϕ maps the state $0 \in \mathcal{X}$ to its own endogenous state (otherwise, AC-State has *already* failed, as desired).

Note that under a uniform exploration policy, where defined,

$$\forall x \in \mathcal{X} \setminus \{0\},\, x' \in \mathcal{X},\, k \in \mathbb{N} \quad \Pr(a_t = L | x_t = x, x_{t+k} = x') = 0.5 \tag{12}$$

that is, it is impossible to determine a_t from the inverse dynamics unless $x_t = 0$ (because for $x_t \neq 0$, the next action a_t does not affect the dynamics). Then it is only useful to distinguish two values of $x \in \mathcal{X} \setminus \{0\}$ if they produce distinct values of $\Pr(a_t = L | x_t = 0, x_k = x)$ for some *particular* $k \leq K$. In particular, **if two states $x, x' \in \mathcal{X} \setminus \{0\}$ cannot each occur exactly k steps after state 0 for some $k \leq K$, then the AC-State loss can be minimized without ϕ distinguishing x and x'.** Because AC-State returns the minimum-range encoder ϕ, it will preferentially return an encoder that maps x and x' to the same endogenous state over the only correct encoder, which maps every x to its own s. Therefore, to show that AC-State will fail for some K, we need only show that there are two states $x, x' \in \mathcal{X} \setminus \{0\}$ which cannot both be reached from state 0 in the same number of steps $\leq K$.

For simplicity and without loss of generality assume that the Ex-BMDP is in state $x = 0$ at timestep $t = 0$. Then the state at timestep k is given by:

$$(k + (q - p)r) \,\%\, q \tag{13}$$

where r is the number of times the action R is selected while the Ex-BMDP is in the state 0, and is at most $\lceil k/p \rceil$ (because the agent can start to take the "short" p-length loop at most this many times in k total steps.) Now, let x, x' be any two states in $\mathcal{X} \setminus \{0\}$ such that

$$(x - x') \equiv (q - p)((q - 1)/2) \pmod{q} \tag{14}$$

(Concretely, if $((q-p)(q-1)/2) \,\%\, q \neq q-1$, we can set $x = q-1$; $x' = q-1-((q-p)(q-1)/2) \,\%\, q$; otherwise we can set $x = q-2$, $x' = q-1$.)

Then, a path of length k ends at x if and only if it contains r R-actions at state 0, where:

$$(k + (q - p)r) \equiv x \pmod{q} \tag{15}$$

and such a path ends at x' if and only if it contains r' R-actions at state 0, where

$$(k + (q - p)r') \equiv x' \pmod{q} \tag{16}$$

Subtracting these two congruences (Eq. 15 and 16) gives:

$$(q - p)(r - r') \equiv x - x' \pmod{q}. \tag{17}$$

Substituting in Eq. 14 gives

$$(q - p)(r - r') \equiv (q - p)((q - 1)/2) \pmod{q}. \tag{18}$$

Because q is prime and $q - p < q$, the quantities $(q - p)$ and q are relatively prime, so we can safely divide off $(q - p)$:

$$r - r' \equiv (q - 1)/2 \pmod{q}. \tag{19}$$

Now either $r - r' > 0$, in which case $r - r'$ is at least $(q-1)/2$, or $r - r' < 0$, in which case $r - r'$ is at most $(q-1)/2 - q = -(q+1)/2$, so $r' - r$ is at least $(q+1)/2$. In either case, because r and r' are non-negative, one of r or r' must be at least $(q-1)/2$. But r and r' must be at most $\lceil k/p \rceil$, as noted above. Therefore, for x and x' to both be reached from state 0 in the same number of steps k, we must have:

$$(q-1)/2 \leq \lceil k/p \rceil \leq (k-1)/p + 1 \tag{20}$$

Then

$$\frac{1}{2}p(q-3) + 1 \leq k \tag{21}$$

Therefore, AC-State will fail if:

$$K \leq \frac{1}{2}p(q-3). \tag{22}$$

Now, we need to frame this bound in terms of the radius D, rather than p and q. First, note that the radius of the MDP is in fact $D = q - 1$ (In particular, the longest distance between any two states is the distance from i to $i-1$, for $0 < i \leq q - p$.)

Secondly, in order to make this bound as tight as possible in terms of D, we would like to make p as close to D as possible. It is known (Polymath, 2014) that there are an infinite number of pairs of primes p, q such that $q - p \leq 246$. Then, in terms of D alone, we have that there are arbitrarily large values of D for which AC-State can fail for all values of K with:

$$K \leq \frac{1}{2}(D-245)(D-2) = \frac{1}{2}D^2 + O(D). \tag{23}$$

If the Twin Prime conjecture holds, then this bound becomes tighter (although not asymptotically): in this case, there are arbitrarily large values of D for which AC-State can fail for all values of K with:

$$K \leq \frac{1}{2}(D-1)(D-2) = \frac{1}{2}D^2 - \frac{3}{2}D + 1 \tag{24}$$

Finally, because we have shown that such MDPs can be constructed for an infinite number of arbitrarily large values of D, we can define $h(D) = \frac{1}{2}(D-245)(D-2)$ for such values of D, and $h(D) = 0$ elsewhere. By the definition of lim sup, we have:

$$\limsup_{D \to \infty} \frac{h(D)}{\frac{1}{2}D^2} = 1 \tag{25}$$

as desired, and thus $h(D) \in \Omega(D^2)$ (by the Hardy-Littlewood definition) as desired. □

While we have shown that we do in fact need to use $K = \Omega(D^2)$ steps in the AC-State method (in environments with aperiodic latent dynamics, where AC-State works at all), it is natural to ask whether $\Omega(D^2)$ steps are still necessary when using the ACDF loss. In the ACDF case, it is *not* enough to show that the multistep inverse loss can conflate two individual states (i.e., to show that two states can have a witness distance quadratic in D, as above). Rather, in order to fail, the method must conflate some *sets* of states such that the resulting dynamics on the combined "states" produced by the encoder are deterministic. In fact, this requirement does not hold for the construction above: empirically, in Section 4 we see that ACDF is able to consistently, successfully infer ϕ even with *one-step* inverse dynamics ($K = 1$) on the version of this MDP with $p = 3$, $q = 5$.[7]

However, we are still able to construct a family of Ex-BMDPs such that ACDF will fail for K smaller than a function $\in \Omega(D^2)$, specifically, a function that goes as $D^2/4$.

[7]Shown in the first numerical experiment at the top of Figure 4. Note that while the bound above for this problem would suggest that AC-State will fail for $K \leq 3$, a tighter analysis of the specific instance shows that it will in fact fail for $K \leq 6$, which is confirmed in the experimental results.

Proposition D.4. *$\forall D$, there exists an Ex-BMDP such that a minimal endogenous latent dynamics has a diameter of D, and such that the ACDF algorithm given by Equation 3 with a uniform exploration policy will return an encoder that does **not** produce a valid endogenous state representation using $K \leq h(D)$ -step inverse dynamics, where $h(D) \in \Omega(D^2)$. Here, we are using the Hardy-Littlewood Ω notation (not to be confused with Knuth's Ω), and specifically $h(D) \sim D^2/4$, in the sense that:*

$$\limsup_{D\to\infty} \frac{h(D)}{\frac{1}{4}D^2} = 1 \tag{26}$$

Proof. Consider two arbitrary primes p, q, with $p < q$, and let these primes define a deterministic Ex-BMDP M with transition function $T(x, a)$ as follows:

- $2q$ states, labeled $\mathcal{X} = \{0, ..., q-2, q-1; 0', 1'..., (q-1)'\}$
- Two actions, $\{L, R\}$
- $T(s, a)$ defined as:
  - $T(0, L) = 1$
  - $T(0, R) = (q - p + 1)'$
  - $T(0', L) = 1'$
  - $T(0, R) = q - p + 1$
  - $\forall x \in \{1, ..., q-1\}$: $T(x, L) = T(x, R) = (x + 1) \,\%\, q$; and $T(x', L) = T(x', R) = ((x + 1) \,\%\, q)'$, where % is the modulo operator.

First, we note that this Ex-BMDP is irreducible and aperiodic. To show that it is irreducible, simply note that we can reach any state from any other state (i.e., any state b can be reached from any state a by simply taking action L $(b - a) \,\%\, q$ times; any state b' can be reached from a' in a similar manner; any b' can be reached from a by first going from a to 0 using L actions as described, using the action R to go from 0 to $(q - p + 1)'$, and then using L actions to go to b'; b can be reached from a' similarly). For aperiodicity, note that the state $0 \in \mathcal{X}$ has a self-loop of length q (by taking L and then any other $q - 1$ actions) and a self-loop of length $2p$ (by taking R and then any other $p - 1$ actions to get to state $0'$, then repeating this process to return to 0). Because p and q are both prime and $q \neq 2$, the g.c.d. of the lengths of these two self-loops is 1, so the Markov chain is aperiodic. Then, by applying Lemma D.2, we know that the *only* valid endogenous latent encoder (up to a relabeling permutations) is the trivial one: $\phi(x) = x$. We can thus regard the states $\mathcal{S} = \mathcal{X}$ as the only "true" endogenous states: any encoder which maps any two states in $\mathcal{X}$ to the same endogenous state must be incorrect.

We will show that there is an alternate encoder, ϕ', such that the dynamics on the inferred "states" produced by ϕ' are deterministic and the multistep inverse loss of ϕ' is minimal if $K \leq h(D)$, where $h(D) \in \Omega(D^2)$. Specifically, $\forall x \in [0, q-1]$, ϕ' maps both x and x' to the same element. To avoid ambiguity, we will refer to this element as x^*. Thus, for example, the element 3^* is the image of the set $\{3, 3'\}$ under ϕ'. Because ϕ' can only output q unique elements, its range is smaller than that of the correct encoder ϕ, and therefore ACDF will produce ϕ' instead of ϕ if both have the same loss.

To show that ϕ' has zero forward dynamics loss, we need to verify that the dynamics on the elements x^* are deterministic. In fact, they are:

- $T(0^*, L) = 1^*$
- $T(0^*, R) = (q - p + 1)^*$
- $\forall x \in \{1, ..., q-1\}$: $T(x^*, L) = T(x^*, R) = ((x + 1) \,\%\, q)^*$;

which happen to be the dynamics of the Ex-BMDP used in the proof of Proposition D.3.

As in the construction used in the proof of Proposition D.3, we note that the first action only affects the dynamics, and hence will only be predictable, if the Ex-BMDP is initially in state 0 or state $0'$. We also note that by the symmetry of the dynamics:

$$\forall x \in \{0, ..., q-1\},\, k \in \mathbb{N} \quad \Pr(a_t = L | x_t = 0, x_{t+k} = x) = \Pr(a_t = L | x_t = 0', x_{t+k} = x') \tag{27}$$

and:

$$\forall x \in \{0, ..., q-1\},\, k \in \mathbb{N} \quad \Pr(a_t = L | x_t = 0', x_{t+k} = x) = \Pr(a_t = L | x_t = 0, x_{t+k} = x') \tag{28}$$

Therefore, up to symmetry, the *only* case in which distinguishing states x and x' (where $x \in \{0, ..., q-1\}$) is necessary to minimize the multistep-inverse loss is when it is needed to distinguish

$$\Pr(a_t = L | x_t = 0, x_{t+k} = x') \tag{29}$$

from

$$\Pr(a_t = L | x_t = 0, x_{t+k} = x) \tag{30}$$

for some *particular* choice of x and k: otherwise ϕ' will be a sufficient encoder to minimize the loss.

For simplicity and without loss of generality assume that the Ex-BMDP is in state $x = 0$ at timestep $t = 0$. Then the state at timestep k is given by:

$$x_k = \begin{cases} (k + (q-p)r) \,\%\, q & \text{if } r \text{ even} \\ ((k + (q-p)r) \,\%\, q)' & \text{if } r \text{ odd} \end{cases} \tag{31}$$

where r is the number of times the action R is selected while the Ex-BMDP is in the state 0 *or* the state $0'$, and is at most $\lceil k/p \rceil$. (because the agent can start to take the "short" p-length path from 0 to $0'$ or vice-versa at most this many times in k total steps.) Then, in order to reach both x and x' in the same number of steps k, we need, for some even r and odd r':

$$(k + (q-p)r) \,\%\, q = (k + (q-p)r') \,\%\, q \tag{32}$$

which implies

$$k + (q-p)r \equiv k + (q-p)r' \pmod{q} \tag{33}$$

Rearranging:

$$0 \equiv (q-p)(r' - r) \pmod{q} \tag{34}$$

Because q is prime and $p < q$, we know that $q - p$ is relatively prime with q, so this equivalence implies:

$$0 \equiv r' - r \pmod{q} \tag{35}$$

Because r' is odd and non-negative, this equivalence implies that $r' \geq q$. But then we have:

$$q \leq r' \leq \lceil k/p \rceil \leq (k-1)/p + 1 \tag{36}$$

Which gives us:

$$(q-1)p + 1 \leq k \tag{37}$$

Therefore, ACDF will fail if:

$$K \leq (q-1)p. \tag{38}$$

Now, we need to frame this bound in terms of the radius D, rather than p and q. First, note that the radius of the MDP is $D = 2q - 1$ (In particular, the longest distance between any two states is the distance from 1 to $(q-p)'$, which requires going "all the way around" the non-primed states to 0, going from 0 to $(q-p+1)'$, and then going "all the way around" the primed states to $(q-p+1)'$.) then in terms of D and the "prime gap" $g := q - p$:

$$K \leq \left(\frac{1}{2}(D-1)\right) \cdot \left(\frac{1}{2}(D+1) - g\right) \tag{39}$$

As discussed in the proof of Proposition D.3, there are an infinite number of pairs of primes with prime gaps bounded by a constant (=246), and under the Twin Primes conjecture, this bound can be made even tighter, with $g = 2$. In either case, we have, for arbitrarily large values of D:

$$K \leq \frac{1}{4}D^2 + O(D). \tag{40}$$

One can then proceed exactly as in Proposition D.3 to prove this proposition. □

E ACDF returns a correct, minimal Encoder

We show both that every function ϕ which minimizes the ACDF loss is a valid endogenous state encoder of the Ex-BMDP, and that there exists a valid, minimal-state endogenous state encoder of the Ex-BMDP which minimizes the ACDF loss. This is sufficient to prove that the minimal-state encoder ϕ' returned Equation 3 will be a valid, minimal-state endogenous state encoder of the Ex-BMDP. Our main result is given as Theorem E.2 below.

To show that a particular ϕ' is a valid endogenous state encoder, we will use the fact that, by the definition of the Ex-BMDP, there must exist *at least one* minimal endogenous state representation, which we will call $s^* \in \mathcal{S}^*$, and a corresponding exogenous state representation, which we will call $e^* \in \mathcal{E}^*$. We assume this encoder follows the assumptions discussed in Section B. Even though it is arbitrarily chosen, for simplicity in our proofs we will call this the "canonical" endogenous/exogenous representation. The "canonical" endogenous encoder ϕ^*, which maps $\mathcal{X}$ to $\mathcal{S}^*$, and "canonical" exogenous encoder ϕ_e^*, which maps $\mathcal{X}$ to $\mathcal{E}^*$, are also defined. We will similarly define the canonical emission distribution $\mathcal{Q}^*$, such that $x_t \sim \mathcal{Q}^*(x|s, e)$, and ϕ^* and ϕ_e^* are inverses of $\mathcal{Q}^*$.

We will also introduce a pair of new objects, which we refer to as the "enhanced exogenous encoder" $\bar{\phi_e}^*(x)$ and "enhanced exogenous state" $\bar{\mathcal{E}}^*$. These are relevant if the dynamics on the canonical endogenous states $\mathcal{S}^*$ are periodic. Specifically, let k be the periodicity of the canonical endogenous transition function T^*, and let $\mathrm{c}(s) \in \mathcal{S}^* \to [0, k-1]$ refer to the cyclic class of the canonical state $s \in \mathcal{S}^*$. (We can fix any labeling of these classes; if T^* is aperiodic, then $\forall s, \mathrm{c}(s) = 0$). Then the "enhanced exogenous state" is defined by concatenating the canonical exogenous state with $\mathrm{c}(\phi^*(x))$; that is, $\bar{\phi_e}^*(x) := (\phi_e^*(x), \mathrm{c}(\phi^*(x)))$. Note that because the cyclic class of s evolves independently of actions, the dynamics of states in $\bar{\mathcal{E}}^*$ is Markovian and exogenous. For a state $e' \in \bar{\mathcal{E}}^*$, we will refer to the two components as $e'[0]$ and $e'[1]$.

We then introduce the following "simulation lemma":

Lemma E.1. *Consider an Ex-BMDP with states $\mathcal{X}$, transition function $\mathcal{T}$, canonical endogenous states $\mathcal{S}^*$ and canonical endogenous states $\mathcal{S}^*$, canonical encoders ϕ^* and ϕ_e^*, and canonical emission distribution $\mathcal{Q}^*$. If some encoder ϕ' exists, such that:*

1. *$\forall x \in \mathcal{X}$, there is a deterministic function which maps from the pair $(\bar{\phi_e^*}(x), \phi'(x))$ to $\phi^*(x)$.*
2. *The dynamics on the encoded latent states produced by $\phi'(x)$ are deterministic. That is, if we fix any $s \in Range(\phi')$ and $a \in \mathcal{A}$, then $\forall x \in \{x|\phi'(x) = s\}$, let $x' \sim \mathcal{T}(x, a)$, then $\phi'(x')$ takes a single deterministic value, a function only of s and a. Additionally, we assume that the range of ϕ' is discrete and finite.*

then ϕ' is a valid endogenous latent encoder for the Ex-BMDP.

Proof. We will show that the Ex-BMDP's transition function $\mathcal{T}$ can be decomposed into endogenous and exogenous parts as defined in Equation 1, such that the endogenous encoder is ϕ', and the exogenous decoder is $\bar{\phi_e}^*(x)$. We will refer to the endogenous/exogenous states under this decomposition as s' and e'. We showed above that the enhanced exogenous state has Markovian dynamics which do not depend on actions; we will refer to these dynamics as $\mathcal{T}_e'$.

By assumption, there is a deterministic function $F(\cdot,\cdot)$, which maps $(\phi'(x), \bar{\phi}^*_e(x))$ to $\phi^*(x)$. There is also, by assumption, a deterministic transition function $T'(\cdot,\cdot)$ which maps $\phi'(x)$ and a to $\phi'(x')$, where x' is any element in the support of $\mathcal{T}(x,a)$.

Then, we can define the emission distribution $\mathcal{Q}'$ as:

$$\forall x \in \mathcal{X}, s' \in \mathcal{S}', e' \in \bar{\mathcal{E}}',\ \Pr_{Q'}(x_t = x | s'_t = s', e'_t = e') := \Pr_{Q^*}(x_t = x | s^*_t = F(s', e'), e^*_t = e'[0]) \tag{41}$$

We now show that:

1. The complete transition function given by:

$$\begin{aligned} &x_{t+1} \sim \mathcal{Q}'(x | s'_{t+1}, e'_{t+1}), \\ &s'_{t+1} = T'(s'_t, a_t), \quad s'_t = \phi'(x_t) \\ &e'_{t+1} \sim \mathcal{T}'_e(e' | e'_t), \quad e'_t = \bar{\phi_e}^*(x), \end{aligned} \tag{42}$$

   is equivalent to the Ex-BMDP transition function $\mathcal{T}$, which by assumption can be expressed as:

$$\begin{aligned} &x_{t+1} \sim \mathcal{Q}^*(x | s^*_{t+1}, e^*_{t+1}), \\ &s^*_{t+1} = T^*(s^*_t, a_t), \quad s^*_t = \phi^*(x_t) \\ &e^*_{t+1} \sim \mathcal{T}^*_e(e^* | e^*_t), \quad e^*_t = \phi^*_e(x), \end{aligned} \tag{43}$$

   To show this fact, is it sufficient to demonstrate that the overall transition probabilities on $\mathcal{X}$ are equal in the two models. That is, we must show that:

$$\Pr_{\mathcal{T}}(x_{t+1} = x | x_t, a_t) = \sum_{e' \in \bar{\mathcal{E}}^*} \Pr_{Q'}(x_{t+1} = x | s'_{t+1} = T'(\phi'(x_t), a_t), e'_{t+1} = e') \cdot \Pr_{\mathcal{T}'_e}(e' | e'_t = \bar{\phi}^*_e(x_t)) \tag{44}$$

.

We can proceed as follows:

$$\sum_{e' \in \bar{\mathcal{E}}^*} \Pr_{Q'}(x_{t+1} = x | s'_{t+1} = T'(\phi'(x_t), a_t), e'_{t+1} = e') \cdot \Pr_{\mathcal{T}'_e}(e' | e'_t = \bar{\phi}^*_e(x_t)) = \quad (45)$$

$$\sum_{e' \in \bar{\mathcal{E}}^*} \Pr_{Q'}(x_{t+1} = x | s'_{t+1} = \phi'(x_{t+1}), e'_{t+1} = e') \cdot \Pr_{\mathcal{T}'_e}(e' | e'_t = \bar{\phi}^*_e(x_t)) = \quad (46)$$

$$\sum_{e' \in \bar{\mathcal{E}}^*} \Pr_{Q^*}(x_{t+1} = x | s^*_{t+1} = F(\phi'(x_{t+1}), e'), e^*_{t+1} = e'[0]) \cdot \Pr_{\mathcal{T}'_e}(e' | e'_t = \bar{\phi}^*_e(x_t)) = \quad (47)$$

$$\sum_{e' \in \bar{\mathcal{E}}^*} \Pr_{Q^*}(x_{t+1} = x | s^*_{t+1} = F(\phi'(x_{t+1}), e'), e^*_{t+1} = e'[0], \mathrm{c}(s^*_{t+1}) = e'[1]) \cdot \Pr_{\mathcal{T}'_e}(e' | e'_t = \bar{\phi}^*_e(x_t)) = \quad (48)$$

$$\sum_{e' \in \bar{\mathcal{E}}^*} \Pr_{Q^*}(x_{t+1} = x | s^*_{t+1} = F(\phi'(x_{t+1}), e'_{t+1}), e^*_{t+1} = e'[0], \mathrm{c}(s^*_{t+1}) = e'[1]) \cdot \Pr_{\mathcal{T}'_e}(e' | e'_t = \bar{\phi}^*_e(x_t)) = \quad (49)$$

$$\sum_{e' \in \bar{\mathcal{E}}^*} \Pr_{Q^*}(x_{t+1} = x | s^*_{t+1} = \phi^*(x_{t+1}), e^*_{t+1} = e'[0], \mathrm{c}(s^*_{t+1}) = e'[1]) \cdot \Pr_{\mathcal{T}'_e}(e' | e'_t = \bar{\phi}^*_e(x_t)) = \quad (50)$$

$$\sum_{e' \in \bar{\mathcal{E}}^*} \Pr_{Q^*}(x_{t+1} = x | s^*_{t+1} = T^*(\phi^*(x_t), a_t), e^*_{t+1} = e'[0]) \cdot \Pr_{\mathcal{T}'_e}(e' | e'_t = \bar{\phi}^*_e(x_t)) = \quad (51)$$

$$\begin{aligned} &\sum_{e' \in \bar{\mathcal{E}}^*} \Pr_{Q^*}(x_{t+1} = x | s^*_{t+1} = T^*(\phi^*(x_t), a_t), e^*_{t+1} = e'[0]) \\ &\quad \cdot \Pr_{\mathcal{T}^*_e}(e'[0] | e^*_t = \phi^*_e(x_t)) \cdot \mathbb{1}_{e'[1] \equiv \mathrm{c}(\phi^*(x_t)) + 1 \pmod k} = \end{aligned} \quad (52)$$

$$\sum_{e'[0] \in \mathcal{E}^*} \Pr_{Q^*}(x_{t+1} = x | s^*_{t+1} = T^*(\phi^*(x_t), a_t), e^*_{t+1} = e'[0]) \cdot \Pr_{\mathcal{T}^*_e}(e'[0] | e^*_t = \phi^*_e(x_t)) = \quad (53)$$

$$\sum_{e^* \in \mathcal{E}^*} \Pr_{Q^*}(x_{t+1} = x | s^*_{t+1} = T^*(\phi^*(x_t), a_t), e^*_{t+1} = e^*) \cdot \Pr_{\mathcal{T}^*_e}(e^* | e^*_t = \phi^*_e(x_t)) = \quad (54)$$

$$\Pr_{\mathcal{T}}(x_{t+1} = x | x_t, a_t) \quad (55)$$

where Eq. 46 follows by the assumption that T' exists and is deterministic (Assumption 2 of the Lemma); Eq. 47 follows by Assumption 1 of the Lemma; Eq. 48 follows from the fact that, if $\Pr_{\mathcal{T}'_e}(e' | e'_t = \bar{\phi}^*_e(x_t))$ is nonzero, then $e'[1]$ must follow $e'_t[1] \pmod k$, therefore $e'[1] = \mathrm{c}(s^*_{t+1})$ for all of these terms; the conditioning is merely making this constraint on e' explicit. Then in Eq. 49 we use the two conditions that $e'[0] = e^*_{t+1}$ and $e'[1] = \mathrm{c}(s^*_{t+1})$ to conclude that $e' = e'_{t+1}$, and substitute using this identity in the other condition. Eq. 50 follows from the definition of $F(\cdot, \cdot)$; and Eq. 51 follows from the definition of the canonical endogenous transition function T^* (we also hide the implicit conditioning/constraint on $e'[1]$, which as discussed above is redundant on all nonzero terms.) Eq. 52 factors the transition function $\mathcal{T}'_e$ into the canonical exogenous transition function $\mathcal{T}^*_e$ and the deterministic procession of the cyclic class of s^*. Note that $e'[1]$ now only appears in the final indicator function term. Furthermore, the indicator is equal to 1 for exactly one of the possible values of $e'[1]$. Therefore, in Eq. 53 we split the sum over $e' \in \bar{\mathcal{E}}^*$ into a sum over $e'[0] \in \mathcal{E}^*$ and a sum over $e'[1] \in [0, k-1]$. The latter sum can be factored to apply only to the indicator function term, and the resulting sum is equal to exactly 1, so it can be eliminated. Eq. 54 is simply a notational change, from $e'[0]$ to e^*. Finally, Eq. 55 holds directly from the definition of the Ex-BMDP in terms of the canonical endogenous and exogenous dynamics.

2. The conditions on Eq. 1 are met:
   - $\mathcal{S}'$, the range of ϕ' is finite, by assumption.
   - The transition function T' is deterministic, by assumption.
   - The transition function $\mathcal{T}'_e$ is Markovian and independent of actions, because it consists of factor $\mathcal{T}^*_e$, which is Markovian and independent of actions, and a cyclic term which evolves deterministically as $n \to (n+1) \,\%\, k$, and is thus also Markovian and independent of actions.

- Q' follows the "block" assumption, because the encoders ϕ' and $\bar{\phi}_e^*$ are both functions, which implies that every value of $x \in \mathcal{X}$ corresponds to a unique (s', e') pair.
- ϕ' and $\bar{\phi}_e^*$ are inverses of Q': in other words, if $x \sim Q'(s', e')$, then $\phi'(x) = s'$ and $\bar{\phi}_e^*(x) = e'$. This statement is only meaningful (i,e., $Q'(s', e')$ is only defined) in cases where s' and e' can co-occur (which is not always the case if $\mathcal{T}_e'$ and T' are both periodic with the same period.) The sets $\mathcal{S}'$ and $\bar{\mathcal{E}}^*$ are defined as the ranges of their respective encoders, so (s', e') can co-occur if and only if there is some $x' \in \mathcal{X}$ such that $s' = \phi'(x')$ and $e' = \bar{\phi}_e^*(x')$. Then we must show that $x \sim Q'(\phi(x'), \bar{\phi}_e^*(x'))$ implies $\phi(x) = \phi(x')$ and $\bar{\phi}_e^*(x) = \bar{\phi}_e^*(x')$. For the first implication, note that, by Eq. 41, we have that $x \sim Q^*(F(\phi(x'), \bar{\phi}_e^*(x')), \bar{\phi}_e^*(x')[0])$, and therefore that $\phi^*(x) = F(\phi(x'), \bar{\phi}_e^*(x'))$ and $\phi_e^*(x) = \bar{\phi}_e^*(x')[0] = \phi_e^*(x')$. Then, by the definition of F, $\phi^*(x) = \phi^*(x')$ and $\phi_e^*(x) = \phi_e^*(x')$. Then, from the dynamics of the canonical latent encoding (specifically, the bounded diameter assumption of $\mathcal{S}^*$ and lack of transient states in $\mathcal{E}^*$), we know that there must exist some x'' such that x'' can transition to both x and x' under the same action a. Then, by the determinism of T', we have that $T'(\phi'(x''), a) = \phi'(x) = \phi'(x')$, so $\phi(x) = \phi(x')$ as desired. Given this result and the fact that $\phi_e^*(x) = \phi_e^*(x')$, it follows by definition that $\bar{\phi}_e^*(x) = \bar{\phi}_e^*(x')$.

□

Given this lemma, we can now prove our main theorem:

Theorem E.2. *Given an Ex-BMDP, under the assumptions listed in Section B, there exists a correct minimal-state endogenous state encoder ϕ^* that will minimize the ACDF loss given in Equation 3 on the Ex-BMDP. Conversely, any encoder ϕ' which minimizes the ACDF loss given in Equation 3 on the Ex-BMDP is a correct endogenous state representation for the Ex-BMDP. As a consequence, the encoder ϕ' which minimizes Equation 3 with the minimum number of output states is a correct minimal-state endogenous representation.*

Proof. Our general approach is as follows:

1. We show that there exists a minimal-state endogenous encoder which *simultaneously* minimizes *both* the multi-step inverse loss term and the latent forward loss term in Equation 3. As a consequence, the encoder also minimizes the overall loss, the sum of these two terms.

2. We show that any ϕ' which achieves the minimum possible value of the *both* terms in the ACDF loss (*independently*) is a correct endogenous-state encoder.

3. Due to item (1) above, we know that it is possible to simultaneously minimize both loss terms. Therefore, *any* ϕ' which achieves the minimum value of the overall loss in Equation 3 must also minimize both loss terms independently. Then we can conclude that any ϕ' which minimizes the overall loss $\mathcal{L}_{\text{ACDF}}$ is a correct endogenous latent state encoder.

4. Finally, because all minimizers of $\mathcal{L}_{\text{ACDF}}$ are correct endogenous state encoders, and we know that at least one of these minimizers is minimal-state among correct encoders, it follows that the minimizers of $\mathcal{L}_{\text{ACDF}}$ which have the fewest numbers of output states are all minimal-state endogenous state encoders.

It show that points (1) and (2) above are correct; points (3) and (4) follow immediately.

(1) A correct, minimal-state encoder minimizes both loss terms.

Let ϕ^* be the correct "canonical" minimal-state endogenous state encoder, as described in Section B. To show that ϕ^* minimizes the multistep inverse loss, we can proceed exactly as in Proposition 5.1 in Lamb et al. (2022).

However, for completeness, we present a version of the proof here. Recall that, by assumption (4) listed in Section B, the behavior policy used to collect data depends only on the endogenous state s^*. Let M be the represent the transition operator of the entire Ex-BMDP on $\mathcal{X}$ under the behavior policy, so that $x_{t+k} \sim M^k x_t$. Similarly, let M_e represent the exogenous-state transition operator under the exogenous transition function $\mathcal{T}_e^*$ and M_s represent the endogenous-state transition operator under the behavioral policy π and the exogenous transition function T^*. Then, by the definition of the Ex-BMDP dynamics, we have that

$$\Pr_{M^k}(x_{t+k} = x|x_t) = \Pr_{Q^*}(x|\phi^*(x), \phi_e^*(x)) \cdot \Pr_{M_e^k}(e_{t+k}^* = \phi_e^*(x)|\phi_e^*(x_t)) \cdot \Pr_{M_s^k}(s_{t+k}^* = \phi^*(x)|\phi^*(x_t)) \tag{56}$$

Additionally:

$$\begin{aligned}&\Pr_{\mathcal{M}^{k-1},\mathcal{T}}(x_{t+k} = x|a_t = a, x_t) = \\ &\Pr_{Q^*}(x|\phi^*(x), \phi_e^*(x)) \cdot \Pr_{M_e^k}(e_{t+k}^* = \phi_e^*(x)|\phi_e^*(x_t)) \cdot \Pr_{M_s^{k-1}}(s_{t+k}^* = \phi^*(x)|s_{t+1}^* = T^*(\phi^*(x_t)))\end{aligned} \tag{57}$$

Now, for any $(x_t, x_{t+k}; k)$ tuple, the multistep-inverse loss in Equation 3 is minimized by the function $f^{\text{opt.}}$ defined as: $f^{\text{opt.}}_{a_t}(x_t, x_{t+k}; k) := \Pr_{\mathcal{D}_{(k)}}(a_t|x_t, x_{t+k})$ where $\mathcal{D}_{(k)}$ is the sampling distribution over which we compute the loss, as defined in Section B; note that the transition function on this distribution is the M defined above. Now, we can write:

$$\begin{aligned}&\Pr_{\mathcal{D}_{(k)}}(a_t = a|x_t, x_{t+k} = x) = \\ &\frac{\Pr_\pi(a_t = a|x_t) \cdot \Pr_{\mathcal{M}^{k-1},\mathcal{T}}(x_{t+k} = x|a_t = a, x_t)}{\Pr_{M^k}(x_{t+k} = x|x_t)} = \\ &\frac{\Pr_\pi(a_t = a|\phi^*(x_t)) \cdot \Pr_{Q^*}(x|\phi^*(x), \phi_e^*(x)) \cdot \Pr_{M_e^k}(e_{t+k}^* = \phi_e^*(x)|\phi_e^*(x_t)) \cdot \Pr_{M_s^{k-1}}(s_{t+k}^* = \phi^*(x)|s_{t+1}^* = T^*(\phi^*(x_t)))}{\Pr_{Q^*}(x|\phi^*(x), \phi_e^*(x)) \cdot \Pr_{M_e^k}(e_{t+k}^* = \phi_e^*(x)|\phi_e^*(x_t)) \cdot \Pr_{M_s^k}(s_{t+k}^* = \phi^*(x)|\phi^*(x_t))} = \\ &\frac{\Pr_\pi(a_t = a|\phi^*(x_t)) \cdot \Pr_{M_s^{k-1}}(s_{t+k}^* = \phi^*(x)|s_{t+1}^* = T^*(\phi^*(x_t)))}{\Pr_{M_s^k}(s_{t+k}^* = \phi^*(x)|\phi^*(x_t))}\end{aligned} \tag{58}$$

Where the we use Equations 56 and 57, and the fact that π depends only on the endogenous state, in the third line. Note that the final line of the equation depends on x and x_t only through $\phi^*(x)$ and $\phi^*(x_t)$ respectively, so $f^{\text{opt.}}$ can be equivalently written as a function only of $\phi^*(x_{t+k})$, $\phi^*(x_t)$, and k. Then the encoder ϕ^* is able to achieve the minimum possible value of the multistep inverse loss term.

We now show that ϕ^* minimizes the latent forward loss as well. Because T^* is deterministic, it follows that, for all transitions (x_t, a_t, x_{t+1}), $\phi^*(x_{t+1})$ is a deterministic function of $\phi^*(x_t)$ and a_t (in particular, $\phi^*(x_{t+1}) = T^*(\phi^*(x_t), a_t)$). Then the forward-prediction loss term will be exactly zero, achieved by setting $g_{s'}(s, a) := \mathbb{1}_{s'=T^*(s,a)}$. This is the minimum possible value for a negative-log loss term.

(2) An encoder ϕ' which minimizes both loss terms is a correct endogenous encoder.

By assumption, the Ex-BMDP must have some canonical endogenous states $\mathcal{S}^*$ and canonical exogenous states $\mathcal{E}^*$, canonical encoders ϕ^* and ϕ_e^*, and canonical emission distribution $\mathcal{Q}^*$. In order to use the above Lemma E.1, we will first show that if ϕ' minimizes the multistep-inverse loss term, then there is a deterministic mapping from $(\bar{\phi_e^*}(x), \phi'(x))$ to $\phi^*(x)$.

We proceed by contradiction. Suppose, for two distinct x, x', we have that $(\bar{\phi}_e^*(x), \phi'(x)) = (\bar{\phi}_e^*(x'), \phi'(x'))$, but $\phi^*(x) \neq \phi^*(x')$. Because $\bar{\phi}_e^*(x)[1] = \bar{\phi}_e^*(x')[1]$, we know that $\phi^*(x)$ and $\phi^*(x')$ must belong to the same cyclic class in $\mathcal{S}^*$; additionally, we have that $\phi_e^*(x) = \phi_e^*(x')$. Let $w = W(\phi^*(x), \phi^*(x'))$. Because $\phi^*(x)$ and $\phi^*(x')$ belong to the same cyclic class, we know by Theorem D.1 that $w \leq D' < \infty$. Let $c \in \mathcal{S}^*$ be the corresponding witness state. By the lack of transient states in $\mathcal{E}^*$, we know that there is some $e^* \in E^*$ such that under the exogenous dynamics $\mathcal{T}_e^*$, $\phi_e^*(x)$ can be reached from e^* in exactly w steps. We can step from (c, e^*) to $(\phi^*(x), \phi_e^*(x))$ in exactly w steps, or from (c, e^*) to $(\phi^*(x'), \phi_e^*(x))$ in exactly w steps. By Lemma B.1, there exists an observation $x_c \in \mathcal{X}$ such that $\phi^*(x_c) = c$ and $\phi_e^*(x_c) = e^*$. Then, by Assumption 5 listed in Section B, we sample $(x_t = x_c, a_t \in \mathcal{A}, x_{t+w} = x)$ with finite probability and $(x_t = x_c, a_t \in \mathcal{A}, x_{t+w} = x')$ with finite probability. By assumption, $\phi'(x) = \phi'(x')$, so

$$f(\phi'(x_c), \phi'(x); w) = f(\phi'(x_c), \phi'(x'); w). \tag{59}$$

Note that the multistep inverse loss term is minimized on *every* tuple $(x_t, x_{t+k}; k)$ by the trivial encoder ϕ_{triv} which maps $\phi_{\text{triv}}(x) := x$. Under this encoder, we can simply define $f_{a_t}^{\text{opt.}}(x_t, x_{t+k}; k) := \Pr_{\mathcal{D}_{(k)}}(a_t | x_t, x_{t+k})$ where $\mathcal{D}_{(k)}$ is the sampling distribution as defined in Section B. Furthermore, for a fixed $(x_t, x_{t+k}; k)$ which occurs with nonzero probability, this minimizer $f^{\text{opt.}}(x_t, x_{t+k}; k)$ is unique: the excess loss conditioned on $(x_t, x_{t+k}; k)$ is exactly the KL-divergence between $\Pr_{\mathcal{D}_{(k)}}(a_t | x_t, x_{t+k})$ and the distribution output by the classifier $f(x_t, x_{t+k}; k)$.

Then, ϕ' can only minimize the loss term if, for some f and all $a \in \mathcal{A}$:

$$\begin{aligned} \Pr_{\mathcal{D}_{(w)}}(a_t = a | x_t = x_c, x_{t+w} = x) = & f_a(\phi'(x_c), \phi'(x); w) = \\ & f_a(\phi'(x_c), \phi'(x'); w) = \Pr_{\mathcal{D}_{(w)}}(a_t = a | x_t = x_c, x_{t+k} = x'). \end{aligned} \tag{60}$$

Where the middle equality is Equation 59.

However, $\Pr_{\mathcal{D}_{(w)}}(a_t = a | x_t = x_c, x_{t+w} = x)$ and $\Pr_{\mathcal{D}_{(w)}}(a_t = a | x_t = x_c, x_{t+w} = x')$ have *disjoint support* over values of $a \in \mathcal{A}$. To see why, suppose that some $a' \in \mathcal{A}$ existed such that both of these probabilities were nonzero. This would imply that, for the canonical endogenous latent state $d \in \mathcal{S}^*$ defined as $d := T^*(c, a')$, it is possible to reach both latent states $\phi^*(x)$ and $\phi^*(x')$ from d in exactly $w - 1$ steps of T^*. But then $W(\phi^*(x), \phi^*(x')) \leq w - 1$, which contradicts the definition of $w := W(\phi^*(x), \phi^*(x'))$.

Then we can conclude that distributions $\Pr_{\mathcal{D}_{(w)}}(a_t = a | x_t = x_c, x_{t+w} = x)$ and $\Pr_{\mathcal{D}_{(w)}}(a_t = a | x_t = x_c, x_{t+w} = x')$ have disjoint support. This leads to a contradiction with Equation 60, which requires the two distributions to be equal. Therefore ϕ' cannot minimize the multistep inverse loss term if, for any two distinct x, x', we have $(\bar{\phi}_e^*(x), \phi'(x)) = (\bar{\phi}_e^*(x'), \phi'(x'))$, but $\phi^*(x) \neq \phi^*(x')$. Therefore for any ϕ' which minimizes the loss term, there must be a deterministic mapping from $(\bar{\phi}_e^*(x), \phi'(x))$ to $\phi^*(x)$.

Secondly, we note that any ϕ' which minimizes the forward dynamics loss term must produce a set of states $\mathcal{S}'$ with a deterministic transition function. To see this, note that this loss term can be zero with *some* encoder (such as the encoder which maps all x to a single latent state), and further note that the loss is *only* zero if $\phi'(x_{t+1})$ is exactly predictable from $\phi'(x_t)$ and a_t; which means that the dynamics are deterministic.

From these two conclusions, that there is a deterministic mapping from $(\bar{\phi}_e^*(x), \phi'(x))$ to $\phi^*(x)$, and that there are deterministic dynamics on the endogenous states produced by ϕ', we can conclude by Lemma E.1 that ϕ' is a valid endogenous encoder of the Ex-BMDP.

□

F Alternative "fixes" to AC-State that do not work

While ACDF has nice theoretical properties, one concern is that, unlike the purely *supervised* action-prediction task of AC-State, the ACDF method's latent-forward prediction task has *moving targets*. In other words, the latent code of the next state $s_t + 1$ will evolve throughout training, complicating the training optimization. Given this difficulty, it would be desirable to come up with a method for provably learning the endogenous encoder of an Ex-BMDP that does *not* rely on moving targets. We considered two initially promising possibilities, but found that both can fail to capture a correct endogenous state representation.

F.1 Full Multi-Step Inverse

F.1.1 What is it?/Why was it Promising?

The multistep inverse method proposed by Lamb et al. (2022) uses $\phi(x_t)$ and $\phi(x_{t+k})$ to predict a_t, that is, the action immediately following x_t. A natural extension of this method would be to model the probability distribution over *all* of the actions on the path between $\phi(x_t)$ and $\phi(x_{t+k})$, that is, $a_t, a_{t+1}., , , , a_{t+k-1}$, based on $\phi(x_t)$ and $\phi(x_{t+k})$ alone. Equivalently, this modeling can be done autoregressively by, for each $k' \in [0, k-1]$, predicting $a_{t+k'}$ given $\phi(x_t)$, $\phi(x_{t+k})$, and $a_t, a_{t+1}., , , , a_{t+k'-1}$. This formulation yields the following modified loss function:

$$\mathcal{L}_{\text{AC-State-Full-Multi}}(\phi_\theta) := \min_f \mathop{\mathbb{E}}_{k\sim\{1,\dots,K\}} \mathop{\mathbb{E}}_{k'\sim\{1,\dots,k-1\}} \mathop{\mathbb{E}}_{(x_t,a_t,\dots,a_{t+k'-1},x_{t+k})} -\log(f_{a_{t+k'}}(\phi_\theta(x_t), \phi_\theta(x_{t+k}), a_t, ..., a_{t+k'-1}; k)) \tag{61}$$

Conceptually, this technique might seem promising because the task of predicting $a_{t+k'}$ from $\phi(x_t)$, $\phi(x_{t+k})$, and $a_t, a_{t+1}., , , , a_{t+k'-1}$ can be accomplished by decomposing the problem into predicting $\phi(x_{t+k'})$ from $\phi(x_t)$ and $a_t, a_{t+1}., , , , a_{t+k'-1}$, and then predicting $a_{t+k'}$ from $\phi(x_{t+k'})$ and $\phi(x_{t+k})$. In other words, it can be accomplished by composing a *latent forward model* with a *multistep inverse model*. Thus, it would seem to require learning a representation similarly rich to the representation learned by ACDF, without dealing with the moving-target issue caused by explicitly learning a forward dynamics model.

Furthermore, the "AC-State-Full-Multi" method successfully learns the encoder for the 5-state periodic Ex-BMDP shown in Figure 3-E in the main text, which AC-State fails on. In particular, the encoder must be able to distinguish states b and c, because $\Pr(a_{t+1} = L | s_t = b, s_{t+3} = b, a_t = L) = 0.5$, while $\Pr(a_{t+1} = L | s_t = c, s_{t+3} = c, a_t = L) = 1$. States d and e can be distinguished similarly.

F.1.2 Why it fails

Unfortunately, while the "AC-State-Full-Multi" prediction task can be accomplished by learning a deterministic latent forward model and a first-action multistep-inverse model, it does not *require* learning these two things. We show a counterexample here. Consider the following Ex-BMDP on the states $\mathcal{X} = \{a, b, c, a', b', c'\}$, with actions $\mathcal{A} = \{A, B, C\}$ and deterministic transition function $T(x, a)$ shown in Figure 10.

The transition function is defined such that, letting x/X and y/Y represent a/A, b/B, or c/C, we have that for *most* inputs, $T(x, Y) = y'$ and $T(x', Y) = y$, with the exceptions that $T(c', B) = c$ and $T(c', C) = b$. This state representation turns out to be minimal (the only correct encoder up to permutation is the trivial one $\phi(x) = x$). To confirm this fact, briefly, note being in $\{a, a'\}$, $\{b, b'\}$, or $\{c, c'\}$ is clearly controllable, and furthermore we must always keep track of whether we are in a primed or non-primed state to determine the effect of actions B and C when we reach $\{c, c'\}$.

However, the "AC-State-Full-Multi" loss will be minimized by an encoder ϕ' that produces only 5 latent states, with a and a' conflated into a single latent state. The resulting endogenous dynamics

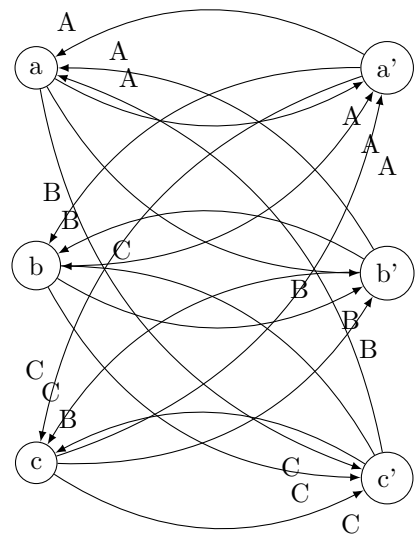


Figure 10: Transitions of $T(x, a)$.

are not deterministic, because starting at the state $\{a, a'\}$ and taking the action B may lead to either the state b or state b', and similarly for for the action C. (See Figure 11).

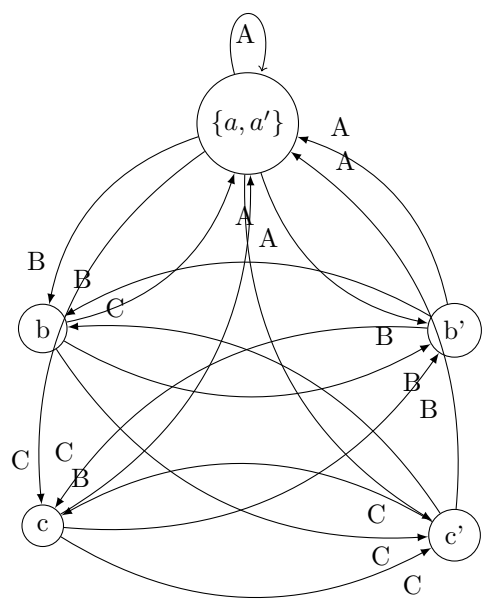


Figure 11: Nondeterministic transitions on $T(\phi'(x), a)$ for the example in Section F.1.2.

To show that this incorrect ϕ' minimizes the "AC-State-Full-Multi' loss, note that, under a uniform policy:

- For any sequence of actions that ends in state a or a', none of the actions can be meaningfully inferred except the *final* action of the sequence, which we know is A with probability 1. Therefore distinguishing whether $x_{t+k} = a$ or $x_{t+k} = a'$ is not necessary to minimize the loss.

- For any sequence of actions that begins with state a or a', but ends in b,b',c or c', one can distinguish if the initial state is a or a' by simply looking at the final state and the parity of the length k of the sequence between them (e.g, if the final state is b, and k is odd, we know the initial state was a' rather than a). Then it is not necessary for the encoder to distinguish between $x_t = a$ and $x_t = a'$.

Therefore, ϕ' is sufficient to minimise the "AC-State-Full-Multi' loss, and is more minimal than the correct endogenous state representation, but is an incorrect endogenous representation. Therefore the "AC-State-Full-Multi" loss fails to discover the true minimal endogenous state representation.

F.2 Artificial Self-Edges/Imprecise k

F.2.1 What is it?/Why was it Promising?

Note that for any bounded-diameter (i.e., irreducible) MDP, adding any self-edges, even a single self-edge, will make the MDP aperiodic. (A self-edge is a transition from a state to itself.) Making this change to the endogenous dynamics of an Ex-BMDP would therefore eliminate the need to use a latent forward model, as in ACDF. In fact, if we add a self-edge to any state, the witness distance between any pair of states will automatically become $\leq D$: if c is the state with the self-edge, then any states a and b can both be reached in exactly $\max(d(c, a), d(c, b))$ steps. Concretely, if a is the further state, then we can reach b in $d(c, a)$ steps by taking the self-edge at c for $d(c, a) - d(c, b)$ timesteps before going to b.

Unfortunately, we cannot simply *alter* the underlying dynamics by adding a self-edge, or many self-edges. However, we might hope to *simulate* such self-edges by randomly duplicating some observations x in the replay buffer, and inserting a "new" action symbol in between the duplicated observations. After learning the state abstraction and dynamics, the self-edges with the "new" action symbol can simply be removed.

A nearly-equivalent idea is to, rather than trying to predict a_t given $\phi(x_t)$, $\phi(x_{t+k})$, and k, instead to predict a_t given $\phi(x_t)$, $\phi(x_{t+k})$, and k', where k' is an *upper bound* on the true value of k (i.e., $k \leq k'$). Note that this formulation is similar to saying that we are given $\phi(x_t)$, $\phi(x_{t+k'})$, and k', but this path of length k' may contain some number $(k' - k)$ of artificially-inserted self-edges. (The only difference is that this second formulation does not allow the *first* action, the predicted action a_t, to be the self-edge, but this distinction is a minor one.) This is an appealing picture, because it directly addresses the core problem with the AC-State method, which is that in order to distinguish two states, they must be *exactly* the same distance k (on some path) from a third state. By allowing some imprecision in k, this method would seem to address this issue.

F.2.2 Why it fails

The problem with this approach is that duplicating an observation effectively "pauses" the *exogenous* state of the Ex-BMDP, not just the endogenous state. This modification makes the exogenous state seem to be "controllable", and therefore may cause the encoder to "leak" information about the exogenous state. Concretely, an incorrect encoder ϕ', with a *larger* range of output "states" than a correct minimal-range encoder ϕ, will have a lower loss than ϕ.

To give an example, consider the Ex-BMDP defined by control-endogenous states $\mathcal{S} = \{a, b, c\}$, exogenous states $\mathcal{E} = \{0, 1\}$, actions $\mathcal{A} = \{L, R\}$ and transitions shown in Figure 12.

The observations $x \in \mathcal{X}$ are defined by simply concatenating the endogenous and exogenous state labels: $\mathcal{X} = \{a0, a1, b0, b1, c0, c1\}$. Let ϕ be a "correct" minimal-state encoder, which maps $a0$ and $a1$ to a, and so on.

The diameter of these dynamics is $D = 2$, and this value is indeed sufficient to learn all endogenous state representations (although the maximum witness distance is in fact $D' = 3$).

(a) Endogenous transitions $T(s, a)$ (b) Exogenous transitions $\mathcal{T}_e(e|e_t)$

Figure 12: An Ex-BMDP with $|\mathcal{S}| = 2$

Let's consider what happens if we use "imprecise k" multi-step inverse, with $k' = 2$. Suppose we have that $x_t = a0$, $x_{t+k} = c1$. If we use the encoder ϕ, our inverse model is only given that $s_t = a$, $s_{t+k} = c$, and $k \in \{1, 2\}$. Then a_t can either be L or R, because one can reach c from a in one step with action L, or in two steps with action R. The exact value of $\Pr(a_t = L|s_t = a, s_{t+k} = c, k \in \{1, 2\})$ will depend on specifics of the implementation: either the distribution from which k is drawn given k', or equivalently, the probability of taking an artificial self-edge (that is, of duplicating any given state on the replay buffer). However, crucially, $\Pr(a_t = L|s_t = a, s_{t+k} = c, k \in \{1, 2\})$ *will not equal 1.* However, if we instead use a *less minimal* encoder ϕ', which distinguishes a0 from a1 and c0 from c1, we can infer by parity that k is in fact equal to 1, from the facts that $x_t = a0$, $x_{t+k} = c1$, and $k \in \{1, 2\}$. We then know that $\Pr(a_t = L) = 1$, because taking L is the only way to reach c from a in a single step. Therefore, the inverse dynamics model will have a smaller loss if the learned encoder ϕ' outputs extra exogenous information. Then this loss function is *not* minimized by the minimal endogenous state encoder.

G Numerical Simulation Experiment Details

G.1 Method

For each Ex-BMDP shown in Figure 4, we perform 50 data-collection runs, and use this data for each value of K and each loss function, to generate 50 trial encoders each. We then determine whether each resulting encoder is either (a) a minimal control-endogenous latent representation, (b) a non-minimal but still correct control-endogenous latent representation, or (c) an incorrect encoder. For each trial run, we:

- Collect two trajectories, each of T timesteps, used as "training" and "validation" datasets. (This setup is similar to the setup of Lamb et al. (2022), where a single trajectory is used for training.)
- Iterate over all possible encoders ϕ. For each ϕ, we:
  - Fit the classifiers f and (g if applicable) on the "training" trajectory, with:

$$f_a(s, s', k) := \frac{\text{Freq. of } (s, a) \xrightarrow{(k)} s' \text{ in Train.}}{\text{Freq. of } s \xrightarrow{(k)} s' \text{ in Train.}}$$

$$g_{s'}(s, a) := \frac{\text{Freq. of } (s, a) \xrightarrow{(1)} s' \text{ in Train.}}{\text{Freq. of } (s, a) \text{ in Train.}}$$

    Note that the states s, s' here are the *encoded* states under ϕ. For both classifiers, if the denominator is zero, we set the distribution as uniform over actions/latent states. If the numerator is zero but the denominator is not, we set the probability as 10^{-7} to avoid infinite losses.

  - Evaluate the expected loss (Equation 2 or 3) on the "validation" trajectory, and record the expected loss as $\mathcal{L}(\phi)$. (The train/validation split is necessary to avoid over-fitting.)

- Find the ϕ with the minimum loss. Among the ϕ's with loss within 0.1% of this overall minimum loss, choose the one with the smallest number of latent states to return.

Note that the number of possible encoders grows extremely quickly with $|\mathcal{X}|$ ($> 10^5$ for 10 states), so we limit our examples to cases with $|\mathcal{X}| = 10$. We also use a uniform random behavioral policy, rather than incorporating planning for exploration as in Lamb et al. (2022), in order to avoid repeatedly searching for the optimal encoder.

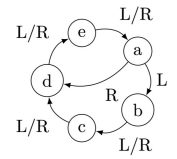
One minor technical caveat is that, for the sake of efficient parallel computation, when measuring the frequency of a transition $(s, a) \xrightarrow{(k)} s'$, we only consider spans $s_t \to s_{t+k}$ for $t \in [0, T - K_{\max} - 1]$, where $K_{\max}$ is the largest K considered in the experiment (e.g., $K_{\max} = 7$ for the first experiment in Figure 4). This wastes a small number (up to $K_{\max} - k$) of possible samples from the trajectory. However, this is negligible compared to the overall length of the trajectory T, and in any case we would not expect this to bias us towards either AC-State or ACDF. (This approximation was also used when computing one-step frequencies for ACDF.)

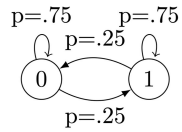
G.2 Additional Results

G.2.1 Success Rates for Correct *Minimal* Encoders

In the previous section, we noted that a returned encoder can either be (a) a minimal control-endogenous latent representation, (b) a non-minimal but still correct control-endogenous latent representation, or (c) an incorrect encoder. In the results shown in the main text, we consider a "success" as either case (a) *or* case (b). That is, we consider all correct encoders as successes, even if they are not minimal-state. Here, in Figure 13, we present the results considering only minimal-state encoders as successful. Interestingly, this only differed at all from the results shown in the main text in two cases (the top and bottom row examples of Figure 4), so we only show results for these cases. For the other two Ex-BMDPs, neither method ever returned a correct but non-minimal encoder.

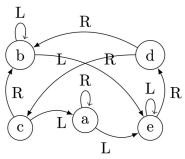
Endogenous Dynamics **T**	Exogenous Noise $\mathcal{T}_e$	AC-State Success Rate	ACDF Success Rate

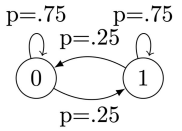




Env. steps:	200	400	800	1600	3200
K=1	0%	0%	0%	0%	0%
K=2	0%	0%	0%	0%	0%
K=3	0%	0%	0%	0%	0%
K=4	0%	0%	0%	0%	0%
K=5	0%	0%	0%	0%	0%
K=6	0%	0%	0%	0%	0%
K=7	72%	86%	94%	92%	96%

Env. steps:	200	400	800	1600	3200
K=1	100%	100%	100%	100%	100%
K=2	100%	100%	100%	100%	100%
K=3	100%	100%	100%	100%	100%
K=4	100%	100%	100%	100%	100%
K=5	100%	100%	100%	100%	100%
K=6	100%	100%	100%	100%	100%
K=7	100%	100%	100%	100%	100%





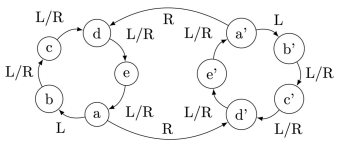
Env. steps:	100	200	400	800	1600
K=1	0%	0%	0%	0%	0%
K=2	46%	48%	34%	30%	30%
K=3	4%	58%	82%	78%	78%
K=4	0%	16%	58%	84%	92%
K=5	0%	0%	42%	86%	100%

Env. steps:	100	200	400	800	1600
K=1	98%	100%	100%	100%	100%
K=2	90%	100%	100%	100%	100%
K=3	64%	100%	100%	100%	100%
K=4	14%	88%	100%	100%	100%
K=5	2%	50%	98%	100%	100%

Figure 13: Results of numerical simulation experiments, where the "Success Rate" includes only cases where the learned encoder is both correct and state-minimal. For the other two examples in Figure 4, no correct but non-minimal encoders were returned, so the results are identical to those shown in the main text.

G.2.2 Complete Results for the Deterministic Ex-BMDP Example (Second Row of Figure 4)

In the example shown on the second row of Figure 4, we tested over a larger range of K than could fit in the figure. The complete results are shown in Figure 14.

Endogenous Dynamics **T**	Exogenous Noise $\mathcal{T}_e$
	(None)

AC-State Success Rate

Env. steps:	1000	2000	4000	8000	16000
K=1	0%	0%	0%	0%	0%
K=2	0%	0%	0%	0%	0%
K=3	0%	0%	0%	0%	0%
K=4	0%	0%	0%	0%	0%
K=5	0%	0%	0%	0%	0%
K=6	0%	0%	0%	0%	0%
K=7	0%	0%	0%	0%	0%
K=8	0%	0%	0%	0%	0%
K=9	0%	0%	0%	0%	0%
K=10	0%	0%	0%	0%	0%
K=11	0%	0%	0%	0%	0%
K=12	0%	0%	0%	0%	0%
K=13	0%	0%	0%	0%	0%
K=14	0%	0%	0%	0%	0%
K=15	0%	0%	0%	0%	0%
K=16	0%	0%	0%	0%	0%
K=17	0%	0%	6%	8%	10%
K=18	0%	0%	2%	2%	4%
K=19	0%	0%	2%	0%	0%
K=20	0%	0%	2%	4%	14%
K=21	0%	0%	2%	4%	10%
K=22	0%	0%	2%	54%	98%
K=23	0%	0%	0%	24%	88%
K=24	0%	0%	0%	16%	82%
K=25	0%	0%	0%	18%	80%
K=26	0%	0%	0%	4%	68%
K=27	0%	0%	0%	4%	62%
K=28	0%	0%	0%	4%	38%
K=29	0%	0%	0%	4%	36%
K=30	0%	0%	0%	2%	30%

ACDF Success Rate

Env. steps:	1000	2000	4000	8000	16000
K=1	2%	14%	0%	0%	0%
K=2	4%	12%	0%	0%	0%
K=3	4%	12%	2%	0%	0%
K=4	4%	10%	2%	0%	0%
K=5	2%	10%	2%	0%	0%
K=6	2%	6%	0%	0%	0%
K=7	2%	2%	0%	0%	0%
K=8	0%	2%	0%	0%	0%
K=9	0%	2%	0%	0%	0%
K=10	0%	2%	0%	0%	0%
K=11	0%	2%	0%	0%	0%
K=12	0%	0%	0%	0%	0%
K=13	0%	12%	22%	64%	96%
K=14	2%	16%	78%	98%	100%
K=15	0%	16%	90%	100%	100%
K=16	0%	22%	96%	100%	100%
K=17	0%	18%	90%	100%	100%
K=18	0%	16%	90%	100%	100%
K=19	0%	12%	88%	100%	100%
K=20	0%	8%	86%	100%	100%
K=21	0%	2%	72%	100%	100%
K=22	0%	0%	68%	100%	100%
K=23	0%	0%	60%	100%	100%
K=24	0%	0%	46%	98%	100%
K=25	0%	0%	42%	98%	100%
K=26	0%	0%	40%	98%	100%
K=27	0%	0%	36%	98%	100%
K=28	0%	0%	32%	98%	100%
K=29	0%	0%	28%	98%	100%
K=30	0%	0%	28%	96%	100%

Figure 14: Full results for the numerical simulation experiment shown in the second row of Figure 4.

H Deep RL Experiment Details

In this section, we describe the deep representation learning experiments we conducted, where the final results are shown in Table 1.

H.1 Environments

We tested on two environments, both shown in Figure 15. The first "baseline" environment is taken from the released code from Lamb et al. (2022), and is similar to the environment described in Section 6.2 of that paper (specifically, it is similar the variant of the environment from that section without "reset" actions). The baseline environment consists of nine copies of a four-room maze, where the controllable agent navigates using four actions (up/down/left/right), in just one of the nine mazes. The other eight mazes have other "agents" in them which take random actions. The observation is a 11×99 image of all nine mazes. For this environment, we start at a random configuration and run for a single trajectory of 5000 steps.

The second environment we tested is designed to be similar in format to the baseline environment, but to specifically highlight the flaws of AC-State. This environment has a periodic transition function, and an action space consisting of two possible actions. To accomplish the periodicity, in this environment, each of the nine identical mazes consists of a track of 40 states, where in most states, the agent simply moves to the next state regardless of action. However, in every fifth state, there is an action which transports the agent to some other position in the track, in the pattern shown in Figure 15. In each of these long-distance jumps, the agent either moves back or forward by a multiple of 10 states. This gives the overall dynamics a periodicity of 10. The maximum witness distance D' of these dynamics is also 10. For data collection, we again collect 5000 transitions; here, we do so as 25 trajectories of length 200 each. In each trajectory, we initialize the state at a random configuration. (The reason for using multiple trajectories is that, because both the ego-agent and the distractor agents have periodic dynamics with the same periodicity, a single episode will not cover the full configuration space.)

H.2 Architecture and Hyperparameters

In general, unless otherwise noted here, we use exactly the architecture and training hyperparameter settings that are default in the released code of Lamb et al. (2022) for gridworld exploration environments.[8]

H.3 Training Schedule and Behavior Policy

We conducted our experiments in an "offline" setting: we collected all 5000 transitions for each each environment under uniformly random actions. We then performed 30,000 training iterations, evaluating after every 5000 steps. When evaluating, we evaluated using the version of ϕ_θ from the previous training steps which achieved the lowest training loss using a rolling average over 20 batches. We exclude the first 3000 training iterations when selecting the lowest-loss previous ϕ_θ.

H.4 Implementation of the forward-dynamics loss

To implement the forward dynamics loss of ACDF (Equation 3), for the dynamics model g we use a LeakyReLU MLP consisting of four layers of sizes $[512 + 10, 1024, 1024, N]$ where the input is the discretized vector output of the vector-quantization output layer of the state encoder, representing the encoded state (a vector of size 512), concatenated with the action label. The output is the discrete *index* of the vector-quantized latent representation code of the next latent state. (The number of codes is a hyperparameter of AC-State which we vary.) Note that because we treat the next-state as a discrete index, we do not backpropagate into the encoder of $\phi_\theta(x_{t+1})$, only into the encoder of $\phi_\theta(x_t)$. This is to mitigate the "moving-target" issue mentioned at the beginning of Appendix F. Also to mitigate this issue, we *only* use the forward prediction loss to update the encoder ϕ_θ every fifth training iteration: at other iterations, we update the parameters of g alone (and separately update ϕ_θ with the multistep-inverse loss).

H.5 Evaluation

In order to capture a real-world usability of the learned representations in planning, we measure success based on the ability to perform open-loop planning to successfully reach a goal.

- We first infer a count-based tabular representation of the deterministic forward dynamics T on the encoded states output by ϕ_θ, by passing each observation of all transitions in the dataset through the final learned ϕ_θ, and inferring that the correct latent transition for a given latent state and action is the one which occurs most often.
- Using this graph, we repeat the following test 1000 times:
  - We first select two random observations x and x', representing a start state and an end state, from the observation space $\mathcal{X}$ if the environment.
  - We record the ground-truth controllable-agent state for both observations, and compute the learned latent states $\phi_\theta(x)$ and $\phi_\theta(x')$. We then use our count-based tabular representation of the learned latent transition function to plan (using Dijkstra's algorithm) a shortest path from $\phi_\theta(x)$ to $\phi_\theta(x')$. This generates a sequence of actions that should (ideally) navigate from x to x'.
  - We execute this sequence of actions from x in an open-loop manner.
  - We count the trial as a success if the ground-truth controllable-agent state after executing these actions is equal to the ground-truth controllable-agent state of x'.

[8] In particular, we used exactly the network architecture for the encoder provided by Lamb et al. (2022)'s released code, for both AC-State and ACDF. We have observed that this architecture design appears to be specifically well-suited to easily learn to filter out the exogenous "distractor" mazes in these multi-maze environments, and may not be more generally applicable. However because we use the same architecture for both ACDF and AC-State, our direct comparison of these methods is valid.

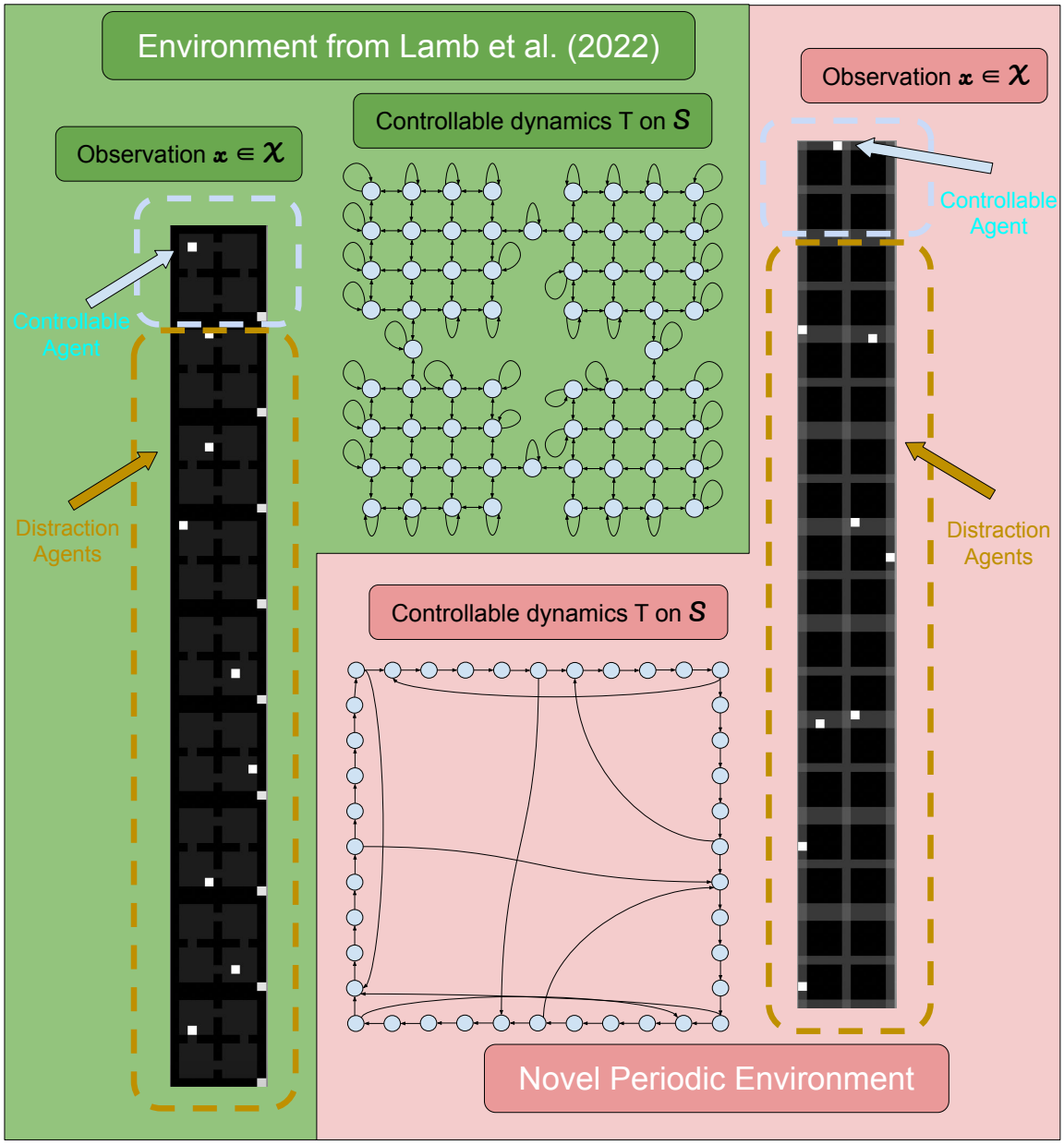


Figure 15: The two high-dimensional environments we tested on. For both environments, we show the image observations $x \in \mathcal{X}$ as well as the states S^* and deterministic state transitions T^* in the controllable latent dynamics (for readability, we do not show action labels). In these environments, the controllable latent dynamics only represent the top maze in the image: the other mazes contain agents which move randomly.

- The overall success rate over the 1000 trials is recorded as a percentage. If it is at least 98% (980/1000), the representation is considered to be successfully learned.

H.6 Results

For both environments and both methods, we ran a hyperparameter sweep over K, the number of steps used in multistep-inverse, and N, the maximum number of latent states (i.e., the size of the codebook of the final vector-quantization layer of the encoder). For each hyperparameter setting, we

ran on 10 seeds. We then evaluated on the "best" version of the hyperparameters for each method for 20 additional seeds. Our final results reflect the rate of successful representation learning at the 98% planning success threshold, at training step 30,000, as described above. Note that the final averages are over the 20 additional seeds alone, to avoid multiple-comparison issues.

To select the "best" hyperparameter setting, we used the following criteria:

- We first used the percentage of successfully-learned representations out of the 10 seeds after 30,000 iterations, based on the 98%-planning success threshold for each seed.
- To break ties, we used the number successfully-learned representations over all 10 seeds and all 6 evaluation times (every 5000 training steps) for each seed. This rewards configurations which learn a successful encoder early, as well as those which may have arrived on successful encoders by chance but then later found a lower-loss encoder for the loss function which was incorrect.
- To break further ties, we used the percentage of *all* open-loop planning trials, over all evaluation times and all seeds, on which the agent succeeded. Note, however, that this is often a misleading statistic, because an incorrect representation of the environment's dynamics might still happen to wind up on the correct place with considerable frequency.

We report the full hyperparameter-sweep results over each of these statistics here (Tables 2-13). We note that while both methods are able to consistently learn the controllable dynamics on the baseline environment, only ACDF consistently learns the dynamics of the periodic environment. By contrast, there is only a single hyperparameter configuration (K=1, N=120) where the lowest-loss encoder under AC-State at the final training iteration was correct, and this was *for only two out of ten random seeds*. (This result is robust to changes in the threshold planning accuracy to be considered a correct representation: these two seeds were the only runs out of the hyperparameter sweep where AC-State succeeded at the final training iteration for success thresholds as low as 75%.)

In the final results shown in the main text, we show that both methods succeeded on the baseline environment on all 20 seeds, while ACDF succeeded on the periodic environment in 19/20 seeds[9], and AC-State only succeeded with one training seed out of 20 on this environment. This is a highly statistically significant difference. Using the Clopper-Pearson method, the 99% CI for the ACDF success rate on this environment is $(.68 - 0.9997)$, while the 99% CI for AC-State is $(.0003 - .32)$. The final open-loop planning accuracies for each of the 20 evaluation seeds for each method and each environment are reported in Tables 14 and 15.

	N=68	N=78	N=88	N=98	N=108
K=3	10.0%	80.0%	90.0%	100.0%	100.0%
K=5	10.0%	80.0%	90.0%	100.0%	100.0%
K=7	0.0%	60.0%	60.0%	90.0%	100.0%

Table 2: AC-State, Baseline Environment, Final-Iteration Complete-Representation Success Rate

	N=68	N=78	N=88	N=98	N=108
K=3	3.3%	58.3%	61.7%	78.3%	80.0%
K=5	3.3%	61.7%	71.7%	76.7%	73.3%
K=7	0.0%	45.0%	46.7%	68.3%	80.0%

Table 3: AC-State, Baseline Environment, All-Iterations Complete-Representation Success Rate

[9]For the remaining one seed, the open-loop navigation task succeeded in 97.5% of trials, falling just barely under the 98% threshold we set to consider the run a success.

	N=68	N=78	N=88	N=98	N=108
K=3	68.8%	85.8%	88.4%	91.5%	91.2%
K=5	66.4%	83.7%	85.4%	88.7%	86.8%
K=7	60.0%	73.7%	78.4%	85.7%	88.4%

Table 4: AC-State, Baseline Environment, All-Iterations Path-Planning Success Rate

	N=68	N=78	N=88	N=98	N=108
K=3	10.0%	90.0%	100.0%	100.0%	100.0%
K=5	40.0%	90.0%	100.0%	100.0%	100.0%
K=7	10.0%	80.0%	100.0%	100.0%	100.0%

Table 5: ACDF, Baseline Environment, Final-Iteration Complete-Representation Success Rate

	N=68	N=78	N=88	N=98	N=108
K=3	6.7%	78.3%	83.3%	81.7%	83.3%
K=5	16.7%	61.7%	76.7%	80.0%	76.7%
K=7	1.7%	48.3%	65.0%	73.3%	76.7%

Table 6: ACDF, Baseline Environment, All-Iterations Complete-Representation Success Rate

	N=68	N=78	N=88	N=98	N=108
K=3	72.1%	87.5%	88.0%	88.3%	88.7%
K=5	72.3%	80.4%	82.6%	84.4%	85.0%
K=7	65.0%	76.1%	77.9%	81.3%	82.4%

Table 7: ACDF, Baseline Environment, All-Iterations Path-Planning Success Rate

	N=40	N=80	N=120
K=1	0.0%	0.0%	20.0%
K=5	0.0%	0.0%	0.0%
K=10	0.0%	0.0%	0.0%

Table 8: AC-State, Periodic Environment, Final-Iteration Complete-Representation Success Rate

	N=40	N=80	N=120
K=1	0.0%	0.0%	11.7%
K=5	0.0%	0.0%	1.7%
K=10	0.0%	0.0%	0.0%

Table 9: AC-State, Periodic Environment, All-Iterations Complete-Representation Success Rate

	N=40	N=80	N=120
K=1	15.3%	28.8%	44.3%
K=5	9.2%	16.6%	26.5%
K=10	4.7%	7.9%	15.6%

Table 10: AC-State, Periodic Environment, All-Iterations Path-Planning Success Rate

	N=40	N=80	N=120
K=1	10.0%	100.0%	100.0%
K=5	0.0%	0.0%	0.0%
K=10	0.0%	0.0%	0.0%

Table 11: ACDF, Periodic Environment, Final-Iteration Complete-Representation Success Rate

	N=40	N=80	N=120
K=1	1.7%	48.3%	46.7%
K=5	0.0%	0.0%	0.0%
K=10	0.0%	0.0%	0.0%

Table 12: ACDF, Periodic Environment, All-Iterations Complete-Representation Success Rate

	N=40	N=80	N=120
K=1	54.8%	65.5%	65.9%
K=5	2.1%	4.2%	4.4%
K=10	2.3%	2.0%	1.9%

Table 13: ACDF, Periodic Environment, All-Iterations Path-Planning Success Rate

AC-State K=1, N=108	100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 99.9%, 99.9%, 99.9%, 99.8%, 99.8%, 99.7%, 99.5%
ACDF K=1, N=108	100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 100.0%, 99.9%, 99.9%, 99.9%, 99.9%, 99.9%, 99.9%, 99.8%

Table 14: Open-loop path planning accuracy for all 20 random seeds for AC-State and ACDF on the baseline environment at final evaluation, after hyperparameter optimization. Values are sorted in descending order of accuracy.

AC-State K=1, N=120	99.6%, 74.6%, 72.0%, 65.4%, 64.5%, 64.4%, 60.2%, 55.0%, 53.8%, 36.3%, 34.8%, 33.9%, 30.8%, 27.8%, 27.5%, 24.4%, 23.8%, 23.6%, 19.8%, 19.0%
ACDF K=1, N=80	100.0%, 99.9%, 99.9%, 99.9%, 99.8%, 99.7%, 99.7%, 99.7%, 99.7%, 99.6%, 99.5%, 99.5%, 99.4%, 99.0%, 99.0%, 99.0%, 98.9%, 98.9%, 98.6%, 97.5%

Table 15: Open-loop path planning accuracy for all 20 random seeds for AC-State and ACDF on the periodic environment at final evaluation, after hyperparameter optimization. Values are sorted in descending order of accuracy.