Offline Action-Free Learning of Ex-BMDPs by Comparing Diverse Datasets

Alexander Levine, Peter Stone, Amy Zhang

Keywords: representation learning, action-free RL, Ex-BMDP, controllable state representations

Summary

While sequential decision-making environments often involve high-dimensional observations, not all features of these observations are relevant for control. In particular, the observation space may capture factors of the environment which are not controllable by the agent, but which add complexity to the observation space. The need to ignore these "noise" features in order to operate in a tractably-small state space poses a challenge for efficient policy learning. Due to the abundance of video data available in many such environments, task-independent representation learning from action-free offline data offers an attractive solution. However, recent work has highlighted theoretical limitations in action-free learning under the Exogenous Block MDP (Ex-BMDP) model, where temporally-correlated noise features are present in the observations. To address these limitations, we identify a realistic setting where representation learning in Ex-BMDPs becomes tractable: when action-free video data from multiple agents with differing policies are available. Concretely, this paper introduces CRAFT (Comparison-based Representations from Action-Free Trajectories), a sample-efficient algorithm leveraging differences in controllable feature dynamics across agents to learn representations. We provide theoretical guarantees for CRAFT's performance and demonstrate its feasibility on a toy example, offering a foundation for practical methods in similar settings.

Contribution(s)

1. We present a provably sample-efficient algorithm, CRAFT, that can learn high-accuracy latent state encoders under the Ex-BMDP model, when provided with two sets of offline observation trajectories, *without action labels*, that are collected by two agents with sufficiently-distinct policies.
   Context: Misra et al. (2024) has shown that efficient representation learning in Ex-BMDPs using a single offline dataset without action labels is in general *not* possible. This work therefore represents to our knowledge the first *positive* theoretical result for this problem. The Ex-BMDP model was introduced by Efroni et al. (2022), who propose a provably sample-efficient algorithm for *online* representation learning in this model. Efroni et al. (2022) assume *near*-deterministic latent-state dynamics, while we make a strict determinism assumption on latent-state dynamics. However, the negative result given by Misra et al. (2024) applies even to the full-determinism variant.
2. We prove the correctness of CRAFT and prove sample-complexity bounds.
   Context: None.
3. We demonstrate the feasibility of CRAFT on a toy problem, and present the results.
   Context: None.

Offline Action-Free Learning of Ex-BMDPs by Comparing Diverse Datasets

Alexander Levine[1], **Peter Stone**[1,2], **Amy Zhang**[1]
alevine0@cs.utexas.edu, pstone@cs.utexas.edu,
amy.zhang@austin.utexas.edu

[1]**The University of Texas at Austin** [2]**Sony AI**

Abstract

While sequential decision-making environments often involve high-dimensional observations, not all features of these observations are relevant for control. In particular, the observation space may capture factors of the environment which are not controllable by the agent, but which add complexity to the observation space. The need to ignore these "noise" features in order to operate in a tractably-small state space poses a challenge for efficient policy learning. Due to the abundance of video data available in many such environments, task-independent representation learning from action-free offline data offers an attractive solution. However, recent work has highlighted theoretical limitations in action-free learning under the Exogenous Block MDP (Ex-BMDP) model, where temporally-correlated noise features are present in the observations. To address these limitations, we identify a realistic setting where representation learning in Ex-BMDPs becomes tractable: when action-free video data from multiple agents with differing policies are available. Concretely, this paper introduces CRAFT (Comparison-based Representations from Action-Free Trajectories), a sample-efficient algorithm leveraging differences in controllable feature dynamics across agents to learn representations. We provide theoretical guarantees for CRAFT's performance and demonstrate its feasibility on a toy example, offering a foundation for practical methods in similar settings.

1 Introduction

Many sequential decision-making settings, such as robotic navigation environments, involve high-dimensional observations with many uncontrollable noise features. In order to efficiently learn policies for many downstream tasks, techniques for task-independent representation learning have been proposed. These techniques learn encoders that map the large observation space into a much smaller set of learned latent states, which can then be used to learn policies for downstream objectives more efficiently than learning from observations directly.

In some such settings, such as social navigation, large amounts of offline *video* data are available, collected either with similar robots or with human agents. In video data, observations are available, but *action* labels are not. Past work has shown that this offline data can be used to learn encodings that can be leveraged for downstream tasks (Ma et al., 2023; Nair et al., 2023; Seo et al., 2022).

However, in recent work, Misra et al. (2024) has shown an important theoretical limitation to this approach: for some important classes of environments, efficient action-free representation learning is not possible. In particular, the Exogenous Block MDP (Ex-BMDP) model (Efroni et al., 2022) describes a class of environments where observations depend both on a deterministic, action-controlled latent state, and a potentially high-dimensional temporally-correlated noise factor, which is action-independent. The goal of representation learning in the Ex-BMDP setting is to learn a mapping from

the observation space to the much-smaller space of action-controlled latent states, while ignoring the noise factor. Misra et al. (2024) show that, even with high coverage over latent states, representation learning from Ex-BMDPs is not possible in general. This property of Ex-BMDPs is in contrast to *Block* MDPs, where the observation noise is not time-correlated, and which Misra et al. (2024) demonstrates *are* amenable to efficient action-free representation learning. At a high level, Misra et al. (2024)'s hardness result stems from the fact that, without action labels, action-controllable features are indistinguishable from uncontrollable features. (For example, if the observations capture both the controllable ego-agent's state and other uncontrollable "background" agents' states, it is ambiguous what state should be encoded in the learned representation.)

In this work, we describe a realistic setting where representation learning for Ex-BMDPs is in fact tractable, and propose a provably sample-efficient algorithm for this setting. Specifically, we consider cases where videos are available of *multiple distinct agents* operating in the same environment. Intuitively, the idea is that *controllable* latent features will *differ* in their dynamics between datasets collected by different agents, while *uncontrollable* features will have *the same* dynamics in the two datasets. Our main result is that, if two agents' policies sufficiently differ at every latent state, then, under assumptions similar to those in Misra et al. (2024) and Efroni et al. (2022), sample-efficient representation learning from offline action-free data collected by the two agents is possible.

To show this fact, we propose a provably sample-efficient algorithm for Ex-BMDP representation learning without action labels, which we call **C**omparison-based **R**epresentations from **A**ction-**F**ree **T**rajectories, or **CRAFT**. CRAFT enjoys a sample complexity that depends only on the size of, and coverage assumptions on, the controllable latent states of the environment, and, logarithmically, on the size of the encoder hypothesis class. The sample complexity has no explicit dependence of the size of the space of exogenous noise. At a high level, CRAFT works by clustering sequential observation-pairs together based on how likely the pairs are to have been observed by each agent.

In this work, we introduce CRAFT, prove its correctness and sample-complexity, and validate its use on a toy example problem. To our knowledge, this is the first work to propose a provably sample-efficient algorithm for action-free offline representation learning in the Ex-BMDP setting. While this work is theoretical in nature due to some restrictive assumptions on the setting (which are inherited from the prior work upon which we build; see discussion in Section 6), we expect that the CRAFT algorithm can inspire practical methods that rely on the same principle of comparing action-free video datasets from diverse agents, in order to extract controllable feature representations.

2 Background

In this section, we define notation, formally introduce the action-free offline Ex-BMDP setting, and state our technical assumptions.

2.1 General Notation

We use $[N]$ to denote the set $\{1, ..., N\}$. For a sequence $x_1, ..., x_N$, we use $x_{i:j}$ to denote the subsequence $x_i, x_{i+1}..., x_j$. For multisets A, B, we use $A \uplus B$ to denote the union of the two multisets, with multiplicities added.

2.2 Ex-BMDP Framework

The Ex-BMDP model describes a class of sequential decision-making environments where an agent's actions only operate on a hidden latent state, while the observations that the agent receives are also functions of a temporally-correlated exogenous "noise" factor, in addition to this controllable latent state. Following Efroni et al. (2022), we consider the *finite horizon* variant of this model,

and also assume that the controllable latent dynamics are *deterministic*.[1] Formally, a (reward-free) Ex-BMDP can be described as a tuple, $\mathcal{M} = \langle H, \mathcal{A}, \mathcal{X}_{1:H}, \mathcal{S}^*_{1:H}, \mathcal{E}_{1:H}, \mathcal{Q}_{1:H}, T_{2:H}, \mathcal{T}^e_{2:H}, s^*_1, P^e_1 \rangle$, where H is the horizon (the number of steps per episode). At each timestep $h \in [H]$, the observation $x_h \in \mathcal{X}_h$ is determined by two latent factors, $s^*_h \in \mathcal{S}^*_h$ and $e_h \in \mathcal{E}_h$. We assume that $\mathcal{S}^*_h$ is finite, while the $\mathcal{E}_h$ and the observation space $\mathcal{X}_h$ may be continuous.

The controllable latent state s^*_h evolves deterministically, depending on the action a_h taken by the agent: $s^*_{h+1} = T_{h+1}(s^*_h, a_h)$, where $T_{h+1} \in \mathcal{S}^*_h \times \mathcal{A} \to \mathcal{S}^*_{h+1}$ is a deterministic function, and $\mathcal{A}$ is the set of possible actions, which we assume to be finite. Note that s^*_1 is a constant. (Each episode starts at the same controllable latent state, so $\mathcal{S}^*_1 = \{s^*_1\}$.)

By contrast, the exogenous (noise) state evolves stochastically as a Markov chain, independent of actions. The initial exogenous state e_1 is sampled from the distribution $P^e_1 \in \mathcal{P}(\mathcal{E}_1)$, and subsequent observations are sampled as $e_{h+1} \sim \mathcal{T}^e_{h+1}(e_h)$, where $\mathcal{T}^e_{h+1} \in \mathcal{E}_h \to \mathcal{P}(\mathcal{E}_{h+1})$. We can refer to the distribution of exogenous states e_h at time h as $P^e_h = \mathcal{T}^e_h(\mathcal{T}^e_{h-1}(...\mathcal{T}^e_2(P^e_1)...))$, and the *joint* distribution of exogenous states e_h and e_{h+1} as $P^e_{h:h+1}$.[2]

The observation x_h is then sampled as $x_h \sim \mathcal{Q}_h(s^*_h, e_h)$, where $\mathcal{Q}_h \in \mathcal{S}^*_h \times \mathcal{E}_h \to \mathcal{P}(\mathcal{X}_h)$ is the *emission function*. Under the Ex-BMDP model, we assume that the latent variables s^*_h and e_h can always be inferred from x_h: that is, $\mathcal{Q}_h$ has a deterministic inverses ϕ^*_h and ϕ^e_h, such that if x_h is sampled from $\mathcal{Q}_h(s^*_h, e_h)$, then $\phi^*_h(x_h) = s^*_h$ and $\phi^e_h(x_h) = e_h$. (This assumption is the *block* assumption referred to in the name "Exogenous Block MDP.")

The agent does not have access to s^*_h, e_h, or the "ground-truth" encoders ϕ^*_h, ϕ^e_h; instead, it only has access to the observations x_h. The goal of representation learning is to learn an encoder $\phi_h : \mathcal{X}_h \to \mathbb{N}$ for each timestep h, that approximates ϕ^*_h, up to label permutation. (We are *not* interested in learning the exogenous encoder ϕ^e_h, because it is assumed that this noise factor is irrelevant for control, and may be very large.)

2.3 Action-Free, Offline Setting

In this work, we consider a setting where the learner has access to multiple sets of offline trajectories collected by different agents, but where only the observations x_h, and *not* the actions a_t, are available. For simplicity, in this work we assume that there are only datasets from two distinct agents, but the proposed method could be straightforwardly generalized to support more agents.

We refer to the two trajectory datasets as τ_A and τ_B. Each trajectory in τ_A (or τ_B) is a sequence of observations $x_{1:H}$. We use $(\tau_A)_{h:h+i}$ to refer to the multiset of tuples of observations $(x_h, x_{h+1}, ..., x_{h+i})$ for each trajectory in τ_A, and use $\tau_A[\{i, i', i''\}]$ to refer to the subset consisting of three *trajectories* (indexed i, i' and i'') in τ_A. We use $\mathcal{D}^*_A(s^*_h, s^*_{h+1})$ to refer to the multiset consisting of all observation pairs $(x_h, x_{h+1}) \in (\tau_A)_{h:h+1}$ such that $\phi^*_h(x_h) = s^*_h$ and $\phi^*_{h+1}(x_{h+1}) = s^*_{h+1}$; and $\mathcal{D}^*(s^*_h, s^*_{h+1}) := \mathcal{D}^*_A(s^*_h, s^*_{h+1}) \uplus \mathcal{D}^*_B(s^*_h, s^*_{h+1})$. We also define $\mathcal{D}^*(s^*_h)$ as the multiset of observations in $x_h \in (\tau_A)_h \uplus (\tau_B)_h$ such that $\phi^*_h(x_h) = s^*_h$.

2.4 Technical Assumptions: Data Collection Method

As in previous works in offline representation learning in the Ex-BMDP setting (Misra et al., 2024; Islam et al., 2023; Levine et al., 2024; Lamb et al., 2023), we assume that the agents' actions a_h are chosen *independently* of the observations $x_{1:h}$, given the controllable latent states $s^*_{1:h}$. In other words, roughly speaking, we assume that the agents used to collect the offline data choose actions based *only* on the controllable latent state, not on the full observation. Misra et al. (2024) justifies this assumption by positing that the offline data are likely collected by expert agents which "would not make decisions based on noise." We discuss this assumption further in Section 6.

[1] Efroni et al. (2022) presents an algorithm for efficient online representation learning of Ex-BMDPs with *near*-deterministic latent dynamics: the controllable latent state deviates from deterministic behavior with frequency $\ll$ one time per episode. See Section 6 for further discussion.

[2] Note that in general, $P^e_{h:h+1} \neq P^e_h \times P^e_{h+1}$.

Beyond sharing this noise-independence assumption, our technical assumptions on the data-collection policies are otherwise significantly weaker that those in Misra et al. (2024). While Misra et al. (2024) assumes that each trajectory is generated by a *Markovian* policy (i.e, that $a_h \sim \pi(s^*_h)$, for some $\pi \in \mathcal{S}^*_h \to \mathcal{P}(\mathcal{A})$), and furthermore that the policies used to generate each trajectory are chosen i.i.d., we make neither such assumption. In other words, we allow for both non-Markovian behavioral policies – for example, we could have a policy in the form $a_h \sim \pi(s^*_{1:h})$ – and non-i.i.d. sampling of behavioral policies between episodes – for example: the data collector could evolve over time between episodes, in order to, for instance, maximize the diversity of visited latent states.

Explicitly, our *only* assumption on the data-collection mechanism is that the process which generates action sequences $a_{1:H}$ – and therefore, equivalently, controllable latent-state sequences $s^*_{1:H}$ – is independent from observation noise over *both entire datasets*. Formally, for a dataset τ that consists of trajectories $x_{1:H}$, let $\phi^*(\tau)$ denote the corresponding set of controllable latent state trajectories $s^*_{1:H}$, and $\phi^e(\tau)$ denote the corresponding set of exogenous state trajectories $e_{1:H}$. Then, our only requirement on the data collection mechanism is that it ensures:

$$\begin{aligned}&\Pr(\tau_A, \tau_B) = \Pr(\phi^*(\tau_A), \phi^*(\tau_B)) \\ &\cdot \Pr_{P^e_1, \mathcal{T}^e}(\phi^e(\tau_A)) \cdot \Pr_{P^e_1, \mathcal{T}^e}(\phi^e(\tau_B)) \cdot \Pr_Q(\tau_A | \phi^*(\tau_A), \phi^e(\tau_A)) \cdot \Pr_Q(\tau_B | \phi^*(\tau_B), \phi^e(\tau_B)).\end{aligned} \tag{1}$$

To see why Equation 1 is a sufficiently strong assumption to allow for useful analysis despite its apparent generality, fix any two latent states $s^*_h, s^*_{h+1} \in \mathcal{S}^*_{h:h+1}$, and consider the multiset $\mathcal{D}^*_A(s^*_h, s^*_{h+1})$ as defined in Section 2.3; also let $n := |\mathcal{D}^*_A(s^*_h, s^*_{h+1})|$. Then the marginal distribution of $\mathcal{D}^*_A(s^*_h, s^*_{h+1})$ can be described as:

$$\mathcal{D}^*_A(s^*_h, s^*_{h+1}) \sim [(\mathcal{Q}(s^*_h, e_h), \mathcal{Q}(s^*_{h+1}, e_{h+1})) | e_h, e_{h+1} \sim P^e_{h:h+1}]^n. \tag{2}$$

We see that $\mathcal{D}^*_A(s^*_h, s^*_{h+1})$ consists of i.i.d. samples from a fixed, policy-independent distribution. Consequently, this property will frequently allow us to use standard concentration bounds in our analysis, while still allowing for a wide class of non-Markovian, non-i.i.d. behavioral policies.

While we do not require the behavioral policies to be Markovian, it will be useful to refer to the "empirical policies" $\pi^{emp.}_A$ and $\pi^{emp.}_B$, defined as:

$$\pi^{emp.}_A(s^*_{h+1} | s^*_h) := \frac{|\mathcal{D}^*_A(s^*_h, s^*_{h+1})|}{\sum_{s' \in \mathcal{S}^*_{h+1}} |\mathcal{D}^*_A(s^*_h, s')|}, \tag{3}$$

and likewise for $\pi^{emp.}_B$. This is the empirical likelihood in the provided data that agent A (respectively, B) chooses an action that results in a transition from s^*_h to s^*_{h+1}.

2.5 Technical Assumptions: Coverage, Policy Diversity, and Realizability

In order to learn accurate latent state encoders $\phi_{1:H}$, we need to ensure adequate coverage over all latent states s_h. For all timesteps h, and all pairs of latent states $(s^*_h, s^*_{h+1}) \in \mathcal{S}^*_h \times \mathcal{S}^*_{h+1}$ such that $s^*_{h+1} = T_{h+1}(s^*_h, a)$ for some action a, we require that

$$\frac{|\mathcal{D}^*(s^*_h, s^*_{h+1})|}{|\tau_A| + |\tau_B|} \geq \nu, \tag{4}$$

for some known lower-bound ν. This coverage assumption is presented in terms of the *actually realized offline datasets* τ_A and τ_B. By contrast, Misra et al. (2024) assumes that trajectories in the offline dataset are sampled i.i.d., and makes coverage assumptions on the *policies* used sample them.

We also require that the two agents, which produced datasets τ_A and τ_B, behaved sufficiently differently so that we can infer the latent dynamics from their differences. In particular, for some known lower bound $\alpha > 0$, we require that, $\forall h \in [H-1], \forall s^*_h \in \mathcal{S}^*_h$, and for any two successor states $s^*_{h+1}, s'^*_{h+1} \in \mathcal{S}^*_{h+1}$, such that s^*_h can transition to either s^*_{h+1} or s'^*_{h+1} under T_{h+1}, we have, either:

$$e^\alpha \cdot \frac{\pi_B^{emp.}(s'^*_{h+1}|s^*_h)}{\pi_B^{emp.}(s^*_{h+1}|s^*_h)} \leq \frac{\pi_A^{emp.}(s'^*_{h+1}|s^*_h)}{\pi_A^{emp.}(s^*_{h+1}|s^*_h)}, \text{ or, } e^\alpha \cdot \frac{\pi_A^{emp.}(s'^*_{h+1}|s^*_h)}{\pi_A^{emp.}(s^*_{h+1}|s^*_h)} \leq \frac{\pi_B^{emp.}(s'^*_{h+1}|s^*_h)}{\pi_B^{emp.}(s^*_{h+1}|s^*_h)}. \tag{5}$$

In other words, the relative likelihood of transitioning to s'^*_{h+1}, versus transitioning to s^*_{h+1}, is different in τ_A and τ_B by a multiplicative factor of at least e^α.

Finally, we also require that the difference in *total* state coverage between τ_A and τ_B for any pair of sequential latent states (s^*_h, s^*_{h+1}) is not *too* extreme. We require that, for a known lower-bound η:

$$\frac{|\mathcal{D}^*_A(s^*_h, s^*_{h+1})|}{|\mathcal{D}^*(s^*_h, s^*_{h+1})|} \geq \eta, \tag{6}$$

and likewise for $\mathcal{D}^*_B(s^*_h, s^*_{h+1})$. Our sample-complexity bound also depends on an additional parameter ν', which does *not* need to be known a priori. This is the minimum single-state coverage ratio:

$$\nu' := \min_{h\in[H], s^*_h \in \mathcal{S}^*_h} \frac{|\mathcal{D}^*(s^*_h)|}{|\tau_A| + |\tau_B|}. \tag{7}$$

Function Approximation Assumptions. We assume access to hypothesis classes of encoder functions $\Phi_{1:H}$, as well as binary classification functions $\mathcal{G}_h \subseteq \mathcal{X}_h \to \{0, 1\}$. We make standard realizability assumptions (in brief, $\phi^*_h \in \Phi_h$, and $\forall s^*_h, s'^*_h, \exists g \in \mathcal{G}_h$ such that g can perfectly distinguish between observations of s^*_h and those of s'^*_h) and assume access to training oracles. See Appendix A for further information about these assumptions. We use $|\Phi|$ to denote $\max_h |\Phi_h|$. We also assume a known upper-bound N_s on $\max_h |\mathcal{S}_h|$; that is, the maximum output range of any encoder in Φ_h.

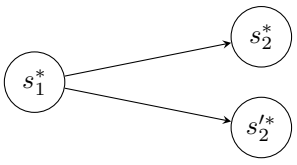
3 Method

In this section, we describe the CRAFT algorithm. First, however, we motivate its design by examining a simpler version of the problem setting and of the algorithm.

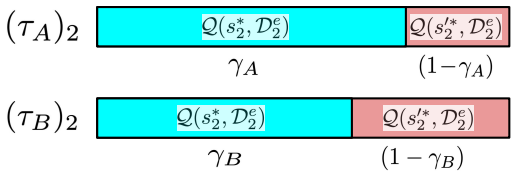
3.1 Intuition: Single-step, Binary Action Case

In this section, we present a toy algorithm for an extremely simplified version of the Ex-BMDP representation learning problem: an explanation of the toy algorithm captures some of the intuitions of CRAFT, while the limitations of the toy algorithm will motivate some of the less-intuitive algorithmic details. We can call this naive, "first draft" form of the CRAFT algorithm "DRAFT."

We first consider the Ex-BMDP model with $H = 2$ and $|\mathcal{A}| = |\mathcal{S}^*_2| = 2$. In this setting, the representation learning problem reduces to the task of learning to distinguish the two latent states s^*_2 and $s'^*_2 \in \mathcal{S}^*_2$ that can occur at $h = 2$. (See Figure 1a.)



(a) Latent-state transition graph.



(b) Composition of the datasets $(\tau_A)_2$ and $(\tau_B)_2$.

Figure 1: Dynamics and composition of the two-step Ex-BMDP example in Section 3.1.

We will also assume that $|\tau_A| = |\tau_B| = m$. Then, $(\tau_A)_1$ and $(\tau_B)_1$ both consist entirely of m i.i.d. samples of the same distribution $\mathcal{Q}(s^*_1, P^e_1)$. The structures of $(\tau_A)_2$ and $(\tau_B)_2$ are (slightly) more

complex. If we let $\gamma_A := \pi_A^{emp.}(s_2^*|s_1^*)$ and $\gamma_B := \pi_B^{emp.}(s_2^*|s_1^*)$, then we see that the dataset $(\tau_A)_2$ consists of $m \cdot \gamma_A$ i.i.d. samples of the distribution $\mathcal{Q}(s_2^*, P_2^e)$ and $m \cdot (1-\gamma_A)$ i.i.d. samples of the distribution $\mathcal{Q}(s_2'^*, P_2^e)$, while $(\tau_B)_2$ consists of $m \cdot \gamma_B$ i.i.d. samples of the distribution $\mathcal{Q}(s_2^*, P_2^e)$ and $m \cdot (1-\gamma_B)$ i.i.d. samples of the distribution $\mathcal{Q}(s_2'^*, P_2^e)$. (See Figure 1b.)

Now, by assumption, the agents generating datasets τ_A and τ_B are not behaviorally identical: this means that $\gamma_A \neq \gamma_B$. Without loss of generality, assume that $\gamma_A > \gamma_B$. Our key insight is that, in the limit as $m \to \infty$, the Bayes optimal classifier to distinguish a sample x_2 selected from $(\tau_A)_2$ from a sample x_2 selected from $(\tau_B)_2$ is in fact (up to label permutation) the latent state encoder $\phi_2^*(x_2)$. Concretely, consider a classifier ϕ_2' trained to minimize the 0-1 classification loss between $(\tau_A)_2$ and $(\tau_B)_2$. In the limit of infinite data, we can define this classification loss function as:

$$\mathcal{L}_{pop}(\phi_2) := \lim_{m\to\infty} \mathop{\mathbb{E}}_{x\in(\tau_A)_2} \phi_2(x) + \mathop{\mathbb{E}}_{x\in(\tau_B)_2} 1 - \phi_2(x), \tag{8}$$

From Equation 8 and the composition of the datasets:

$$\begin{aligned}
\mathcal{L}_{pop}(\phi_2) &= \gamma_A \mathop{\mathbb{E}}_{x\sim\mathcal{Q}(s_2^*,P_2^e)} \phi_2(x) + (1-\gamma_A) \mathop{\mathbb{E}}_{x\sim\mathcal{Q}(s_2'^*,P_2^e)} \phi_2(x) \\
&\quad + \gamma_B \mathop{\mathbb{E}}_{x\sim\mathcal{Q}(s_2^*,P_2^e)} (1-\phi_2(x)) + (1-\gamma_B) \mathop{\mathbb{E}}_{x\sim\mathcal{Q}(s_2'^*,P_2^e)} (1-\phi_2(x)) \\
&= (\gamma_A - \gamma_B)[\mathop{\mathbb{E}}_{x\sim\mathcal{Q}(s_2^*,P_2^e)} \phi_2(x) + \mathop{\mathbb{E}}_{x\sim\mathcal{Q}(s_2'^*,P_2^e)} 1 - \phi_2(x)] + C \\
&= -(\gamma_A - \gamma_B)(\Pr(\phi_2(x)=0|\phi^*(x)=s_2^*) + \Pr(\phi_2(x)=1|\phi^*(x)=s_2'^*)) + C
\end{aligned} \tag{9}$$

where C is independent of ϕ_2. Under the mapping $(0 \to s_2^*, 1 \to s_2'^*)$, we see that $\mathcal{L}_{pop}(\phi_2)$ scales linearly with the rate that ϕ_2 produces incorrect encodings, with a δ-increase in $\mathcal{L}_{pop}$ corresponding to an $\mathcal{O}((\gamma_A - \gamma_b)\cdot\delta)$ increase in encoder failure. In particular, $\mathcal{L}_{pop}$ is uniquely minimized by the ground-truth encoder ϕ_2^*. We now examine how this simple algorithm functions with finite datasets:

Algorithm 1 "DRAFT" algorithm for $H = 2$ Ex-BMDPs.

Require: Datasets τ_A, τ_B with $H = 2$, hypothesis class $\Phi_2 \in \mathcal{X}_2 \to \{0, 1\}$.

Let $\phi_1' := \mathcal{X}_1 \to 0$, and $\phi_2' := \arg\min_{\phi_2\in\Phi_2} \mathcal{L}(\phi_2)$, where:

$$\mathcal{L}(\phi_2) := \frac{1}{|\tau_A|} \sum_{x\in(\tau_A)_2} \phi_2(x) + \frac{1}{|\tau_B|} \sum_{x\in(\tau_B)_2} 1 - \phi_2(x). \tag{10}$$

Return: ϕ_1', ϕ_2'

With finite m, our main concern is overfitting: if Φ_2 is large enough such that some $\phi_2' \in \Phi_2$ can perfectly distinguish the samples that happen to fall into $(\tau_A)_2$ from those in $(\tau_B)_2$, then this ϕ_2' will attain a lower empirical loss than ϕ_2^*, while being bad at distinguishing $\mathcal{Q}(s_2^*, P_2^e)$ from $\mathcal{Q}(s_2'^*, P_2^e)$ in general. However, as long as $|\Phi_2|$ is controlled, we can use standard concentration inequalities to limit this overfitting. In particular,

$$|\mathcal{L}(\phi_2) - \mathcal{L}_{pop}(\phi_2)| \approx \mathcal{O}(1/\sqrt{m}). \tag{11}$$

By combining Equations 11 and 9, we can determine how quickly $\phi_2' = \arg\min \mathcal{L}(\cdot)$ will approach $\phi_2^* = \arg\min \mathcal{L}_{pop}(\cdot)$ as m increases, in terms of *accuracy as a latent state encoder*. To ensure ϕ_2' approximates ϕ_2^* with a failure rate of at most ϵ, we need $m \approx \mathcal{O}\left(\frac{1}{(\gamma_A-\gamma_b)^2\epsilon^2}\right)$ samples. Intuitively, the smaller the difference between behavior policies of the two agents $(\gamma_A - \gamma_B)$, the more samples are required to attain a given accuracy of encoder.

3.1.1 "DRAFT" doesn't generalize easily to the long-horizon setting

A naive first attempt to extend "DRAFT" to the $H > 2$ case might be to apply it *recursively*. That is, once the two distinct latent states at $h = 2$ can be decoded, we can extract from $(\tau)_A$ and $(\tau)_B$

the trajectories which contain (say) s_2^*, and then run DRAFT again on these samples, to obtain an encoder that can separate the two states into which s_2^* can transition.[3] We can then repeat this procedure for $s_2'^*$. If $s_2'^*$ and s_2^* both transition to the same latent state (say $s_3'^*$), we can easily detect this situation by attempting to learn binary classifiers between the observations of each state that succeeds s_2^* and each state that succeeds $s_2'^*$: if it is impossible distinguish these observations better than by random chance, then the two successor states are the same:

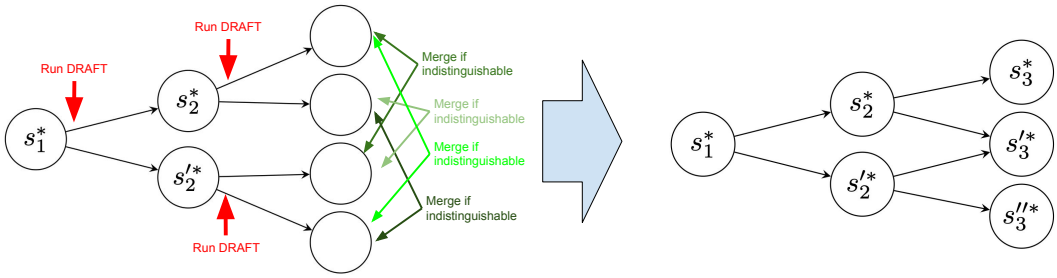


Figure 2: Illustration of recursive use of the "DRAFT" algorithm.

However, it turns out that it is difficult (and may be impossible) to prove an efficient sample-complexity bound for this recursive algorithm. This is for two reasons:

1. After the first timestep, the input datasets to subsequent applications of DRAFT are corrupted by mis-classified samples, such that the datasets are no longer mixtures of i.i.d. samples from distributions $\mathcal{Q}(s^*, P_h^e)$ for multiple values of $s^* \in \mathcal{S}_h^*$.
2. DRAFT is highly sensitive to small changes in its input dataset.

To see (1), note that the encoder ϕ_2' returned by the first application of DRAFT will misclassify some $\mathcal{O}(\sqrt{m}/(\gamma_A - \gamma_B))$ samples. Moreover, these misclassified samples will *not* be chosen uniformly: the encoder ϕ_2' may rely on some spurious features of the observations x_2, which depend on e_2, to classify these observations. Consequently, because e_3 *also* depends on e_2, the exogenous noise distributions of the observations x_3 of state $s_3'^*$ (for instance, in the dynamics example in Figure 2) present in the datasets for the recursive DRAFT instances associated with s_2^* and $s_2'^*$ will differ from each other, and each will differ from $\mathcal{Q}(s_3'^*, P_h^e)$, in a way that depends on the choice of ϕ_2'. Moreover, because ϕ_2' is trained to distinguish τ_A from τ_B, this distribution shift may have different effects on the distributions of observations from the two agents.

For (2), consider the (single step) DRAFT algorithm with some small number $\epsilon_{bad} \cdot m$ of samples from $\mathcal{X}_2$ arbitrarily introduced to the datasets $(\tau_A)_2$ and $(\tau_B)_2$. For concreteness, we will replace some of the samples in $(\tau_A)_2$ for which $\phi_2^*(x) = s_2'^*$ with "bad" samples for which it *still* holds that $\phi_2^*(x'_{bad}) = s_2'^*$, but which are not drawn i.i.d. from $\mathcal{Q}(s_2'^*, P_2^e)$. Instead, we assume that these samples belong to some part of the support of $\mathcal{Q}(s_2'^*, P_2^e)$ which is typically sampled with negligible probability; we can call their distribution $\mathcal{D}'_{bad}$. Similarly, we replace $\epsilon_{bad} \cdot m$ samples of $(\tau_B)_2$ for which $\phi_2^*(x) = s_2^*$ with "bad" samples for which $\phi_2^*(x_{bad}) = s_2^*$, but which are drawn from $\mathcal{D}_{bad}$. We consider the infinite-dataset limit. From Equation 8 and the composition of the datasets:

$$\begin{aligned}
\mathcal{L}_{pop}(\phi_2) &= \gamma_A \mathop{\mathbb{E}}_{x\sim\mathcal{Q}(s_2^*,P_2^e)} \phi_2(x) + (1-\gamma_A-\epsilon_{bad}) \mathop{\mathbb{E}}_{x\sim\mathcal{Q}(s_2'^*,P_2^e)} \phi_2(x) + \epsilon_{bad} \mathop{\mathbb{E}}_{x\sim\mathcal{D}'_{bad}} (\phi_2(x)) \\
&+ (1-\gamma_B) \mathop{\mathbb{E}}_{x\sim\mathcal{Q}(s_2'^*,P_2^e)} (1-\phi_2(x)) + (\gamma_B-\epsilon_{bad}) \mathop{\mathbb{E}}_{x\sim\mathcal{Q}(s_2^*,P_2^e)} (1-\phi_2(x)) + \epsilon_{bad} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{bad}} (1-\phi_2(x)) \\
&= -(\gamma_A-\gamma_B+\epsilon_{bad})\big(\Pr(\phi_2(x)=0|x\sim\mathcal{Q}(s_2^*,P_2^e)) + \Pr(\phi_2(x)=1|x\sim\mathcal{Q}(s_2'^*,P_2^e))\big) \\
&\quad - \epsilon_{bad}\big(\Pr(\phi_2(x)=1|x\sim\mathcal{D}_{bad}) + \Pr(\phi_2(x)=0|x\sim\mathcal{D}'_{bad})\big) + C
\end{aligned}$$

[3] Here, we are continuing to assume $|\mathcal{A}| = 2$, and that the two actions have different effects from each other in each state.

Note that the ground-truth encoder ϕ_2^* has loss $\mathcal{L}_{pop}(\phi_2^*) = -(\gamma_A - \gamma_B + \epsilon_{bad}) + C$. However, we can construct an encoder ϕ_2' that incorrectly encodes all samples in $\mathcal{D}_{bad}$ as belonging to $s_2'^*$ (i.e., returns 1 on these samples), and incorrectly encodes all samples in $\mathcal{D}'_{bad}$ as belonging to s_2^*. Furthermore, we can construct this ϕ_2' to also have an accuracy of only $1 - \epsilon_{bad}/(\gamma_A - \gamma_B + \epsilon_{bad})$ on the samples in the "natural" distributions $\mathcal{Q}(s_2^*, P_2^e)$ and $\mathcal{Q}(s_2'^*, P_2^e)$. Surprisingly, by the above expression for $\mathcal{L}_{pop}$, we see that this less-accurate encoder *also* has loss $\mathcal{L}_{pop}(\phi_2') = -(\gamma_A - \gamma_B + \epsilon_{bad}) + C$. Therefore, if this encoder ϕ_2' is included in the hypothesis class Φ_2, then the ERM step of Algorithm 1 may just as easily return ϕ_2' rather than ϕ_2^*. (Furthermore, the realizability assumption does not guarantee that any lower-loss "third option" encoders exist.) Consequently, the misclassification rate can "blow up" from ϵ_{bad} to $\epsilon_{bad}/(\gamma_A - \gamma_B + \epsilon_{bad})$, that is, by a *multiplicative* factor of $1/(\gamma_A - \gamma_B + \epsilon_{bad})$ – even before accounting for finite datasets.

We can now see that DRAFT both (1) can make non-uniformly-distributed encoding errors, and (2) given as input a dataset with (non-uniform) errors, can produce output "next-state" datasets with a *multiplicatively-increased* error rate. Therefore, it seems difficult to derive a sample-complexity analysis of "recursive DRAFT" that does not require a number of samples exponential in the Ex-BMDP horizon H.[4] In the next section, we present our CRAFT algorithm, which is intentionally designed to solve the multiple-agent action-free Ex-BMDP representation learning problem while *avoiding* recursively training state classifiers on datasets derived from the output of previous-timestep state classifiers. We then can sidestep the issues with "Recursive DRAFT" shown here.

Note that, for Ex-BMDPs with deterministic latent dynamics, the issues with recursion seen here are unique to the offline, action-free setting. In the online setting, as in Efroni et al. (2022), once the dynamics up to timestep h have been learned, "fresh" samples of any given latent state s_h^* can then be constructed via closed-loop planning: there are no issues with compounding error.[5] The action-free offline setting thus presents a new set of issues requiring a novel algorithmic solution.

3.2 CRAFT: High-Level Description of Method

Here, we give a high-level overview of the CRAFT algorithm. The complete algorithm is presented as Algorithm 2 in Appendix C. See also Figure 3 for a pictorial overview of the approach.

In CRAFT, we initially treat each trajectory as a sequence of observation *pairs*: (x_1, x_2), (x_2, x_3), ... (x_{H-1}, x_H). (See Figure 3-a.) For each timestep-pair $(h, h+1)$, we train a model f_h to predict, given a sample (x_h, x_{h+1}), whether the pair was collected by agent A or agent B: that is, whether (x_h, x_{h+1}) was selected from $(\tau_A)_{h,h+1}$ or $(\tau_B)_{h,h+1}$. However, unlike in "DRAFT", we do not treat this problem as hard binary classification. Instead, we train $f_h(x_h, x_{h+1})$ to predict the *log-odds ratio* between the two possibilities, for a given (x_h, x_{h+1}). That is, we train f_h to predict

$$\ln\left(\frac{\Pr[(x_h, x_{h+1}) \in (\tau_A)_{h,h+1} | (x_h, x_{h+1}) \in (\tau_A \uplus \tau_B)_{h,h+1}]}{\Pr[(x_h, x_{h+1}) \in (\tau_B)_{h,h+1} | (x_h, x_{h+1}) \in (\tau_A \uplus \tau_B)_{h,h+1}]}\right). \tag{12}$$

To accomplish this task, we can train f_h to minimize the following loss function:

$$\mathcal{L}(f_h) := \sum_{(x_h, x_{h+1}) \in (\tau_A)_{h,h+1}} \ln(1 + e^{-f(x_h, x_{h+1})}) + \sum_{(x_h, x_{h+1}) \in (\tau_B)_{h,h+1}} \ln(1 + e^{f(x_h, x_{h+1})}). \tag{13}$$

Note that in the limit of infinite data, the f_h that minimizes this loss will return

$$f_h^*(x_h, x_{h+1}) \to \ln\left(\frac{|\mathcal{D}_A^*(\phi_h^*(x_h), \phi_{h+1}^*(x_{h+1}))|}{|\mathcal{D}_B^*(\phi_h^*(x_h), \phi_{h+1}^*(x_{h+1}))|}\right). \tag{14}$$

[4] We are not claiming that "Recursive DRAFT" *actually does* require exponential samples in H, simply that there are clear obstacles to proving that it *does not.*

[5] Even with *offline* data, if action labels are available and the latent dynamics up to timestep h are known perfectly (which is achievable if the latent dynamics are deterministic), then "error-free" datasets can still be constructed for timestep $h+1$.

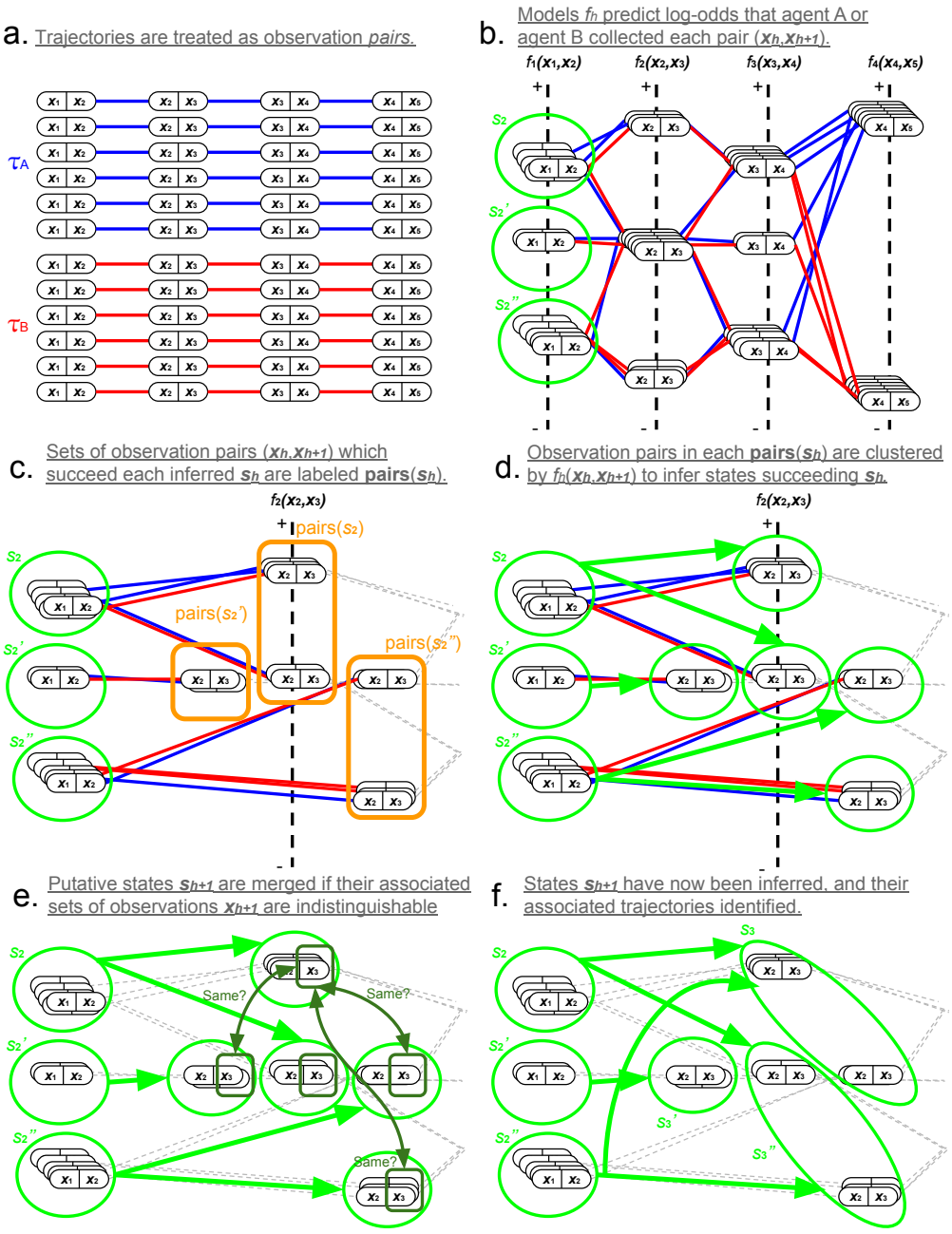


Figure 3: Schematic of the CRAFT algorithm. See text of Section 3.2.

Consequently (for sufficiently-large datasets) we expect the values of $f_h(x_h, x_{h+1})$ of all observation-pairs (x_h, x_{h+1}) corresponding to the same latent-state pair $(\phi_h^*(x_h), \phi_{h+1}^*(x_{h+1})) = (s_h^*, s_{h+1}^*)$ to "cluster together" around the same value (See Figure 3-b): this effect can be quantified using standard concentration arguments. Note that the training of models f_h and resulting "clustering" of observation-pairs can be carried out *simultaneously and independently* for all time-steps h: there is no "recursion" here, and so each model f_h is trained on an "untainted" dataset.

Side note on realizability and discretization: To ensure that an f_h can be found that minimizes Equation 13, we need f_h to be chosen from a sufficiently-expressive hypothesis class $\mathcal{F}_h$. We can construct such an $\mathcal{F}_h$ as $\Phi_h \times \Phi_{h+1} \times (N_s^2 \to \mathbb{R})$: by the realizability assumptions on Φ_h and

Φ_{h+1}, we are ensured that this class contains the optimal predictor in Equation 14. However, the $(N_s^2 \to \mathbb{R})$ component of this hypothesis class makes it non-finite. In order to allow for a simple finite-hypothesis analysis, we instead construct $\mathcal{F}_h$ as $\Phi_h \times \Phi_{h+1} \times (N_s^2 \to \Xi)$, where Ξ is a discrete space (roughly, every $(\alpha/4)$-th interval on a range determined by η). It turns out that this discretization still ensures that a function "close enough" to f_h^* will always exist, and additionally greatly simplifies the identification of "clusters" in the output distribution of f_h on finite data.

While we expect the values of $f_h(x_h, x_{h+1})$ to cluster for sets of observations-pairs with the *same* latent-state-pair, it does not immediately follow that the values of $f_h(x_h, x_{h+1})$ and $f_h(x'_h, x'_{h+1})$ will *differ* if $(s_h, s_{h+1}) \neq (s'_h, s'_{h+1})$. In fact, this is not true in general: two distinct "clusters" may overlap entirely. This is where the assumption given in Equation 5 becomes useful: from Equation 5 and algebra, we can see that for any fixed s_h^* and distinct $s_{h+1}^*, s_{h+1}'^*$ which can both follow s_h^* that:

$$\left| \ln\left(\frac{|\mathcal{D}_A^*(s_h^*, s_{h+1}^*)|}{|\mathcal{D}_B^*(s_h^*, s_{h+1}^*)|}\right) - \ln\left(\frac{|\mathcal{D}_A^*(s_h^*, s_{h+1}'^*)|}{|\mathcal{D}_B^*(s_h^*, s_{h+1}'^*)|}\right) \right| \geq \alpha. \tag{15}$$

In other words, the "clusters" associated with two pairs (s_h^*, s_{h+1}^*) and $(s_h'^*, s_{h+1}'^*)$ are guaranteed to be distinct if $s_h^* = s_h'^*$. One consequence is that the observation-pairs (x_1, x_2) associated with each possible latent-state pair (s_1^*, s_2^*) must form distinct, well-separated clusters (because these pairs all share the same initial latent state s_1^*). Therefore, datasets of observations associated with each latent state s_2 can be immediately identified (as shown in green in Figure 3-b; note that we omit the asterisk, to indicate that these are *inferred*, rather than *ground-truth*, latent states.).

CRAFT then continues "recursively": once the trajectories which contain a particular $s'_2 \in \mathcal{S}$ are known, we can then examine the spectrum of values of $f_2(x_2, x_3)$ for *only* observation pairs (x_2, x_3) from this *subset* of trajectories (referred to in Algorithm 2 as pairs(s'_2)). Because these observation-pairs all (up to an error factor) share the same initial state $s_2'^*$, we expect to see well-separated clusters for each latent state which can succeed s'_2. Note that we *do not* retrain f_2 on only these samples in pairs(s'_2). Therefore, any errors (missing or extra trajectories) in the construction of pairs(s'_2) can only substantially affect the outcome of this step by compromising CRAFT's ability to recognize distinct clusters in the *precomputed values* of $f_2(x_2, x_3)$. Due to the discretization of the range of $\mathcal{F}_h$, this "cluster identification" is robust to even adversarial errors affecting a bounded number of trajectories. The total number of misclassified trajectories then grows only linearly with H. (In Figure 3-c, we show the spectrum of values of $f_2(x_2, x_3)$ for each subset pairs(s_2), pairs(s'_2), and pairs(s''_2); Figure 3-d shows the result of the cluster identification: the observation-pairs (x_2, x_3) corresponding to each state in $\mathcal{S}_3$ which can succeed each of s_2, s'_2, and s''_2 have been identified).

Once we identify each latent state that can succeed each $s'_2 \in \mathcal{S}_2$ *individually*, we now determine whether or not any of these successor states to distinct states s_2, s'_2 are in fact *the same* latent state $s_3 \in \mathcal{S}_3$. This can be accomplished easily, by attempting to learn binary classifiers between the observations x_3 which are part of the observation-pair sets. If these observations are indistinguishable, then the sets of observation-pairs represent the same latent state; if they are perfectly distinguishable, then they represent different latent states. Figure 3-e illustrates this process. Note that while there may be errors in these observation sets, each binary-classifier training ultimately produces a boolean result (either the sets are distinguishable, or they are not) with a substantial allowance for error in the input sets: there is (with high probability) no accumulation of errors due to this process.

Finally, the observations corresponding to each unique latent state $\mathcal{S}_3$ have been identified. (See Figure 3-f). We can then continue to timestep $h = 4$, and so on. As mentioned above, both the cluster-identification and state-merging processes are robust to bounded errors in their input data, so the total number of misclassified states grows only linearly in H. As a final step, the encoders ϕ'_h are trained on the assembled datasets for each timestep h.

3.3 Guarantees

We prove the following polynomial sample-complexity guarantee for CRAFT in Appendix C:

Theorem 3.1. *Assume that CRAFT (Algorithm 2 in the Appendix) is given datasets τ_A and τ_B such that the assumptions given in Equations 1, 4,5, and 6 all hold. Then there exists an*

$$f\left(H, |\Phi|, N_s, \frac{1}{\delta}, \frac{1}{\epsilon_0}, \frac{1}{\nu}, \frac{1}{\nu'}, \frac{1}{\eta}, \frac{1}{\alpha}\right) \in \mathcal{O}^*\left(\frac{H^2(\ln(|\Phi|/\delta) + N_s^2)}{\nu\eta^2\alpha^4} \cdot \max\left(\frac{1}{\nu^2}, \frac{1}{\epsilon_0^2\nu'^2}\right)\right), \quad (16)$$

where $\mathcal{O}^*(f(x)) := \mathcal{O}(f(x)\log^k(f(x)))$, such that for any given $\delta, \epsilon_0 \geq 0$, if $\forall s_h^*, s_{h+1}^*$ such that s_h^* can transition to s_{h+1}^*, $|\mathcal{D}^*(s_h^*, s_{h+1}^*)| \geq f\left(H, |\Phi|, N_s, \frac{1}{\delta}, \frac{1}{\epsilon_0}, \frac{1}{\nu}, \frac{1}{\nu'}, \frac{1}{\eta}, \frac{1}{\alpha}\right)$, then, with probability at least $1-\delta$, the encoders ϕ_h' returned by the algorithm will each have accuracy on at least $1-\epsilon_0$, in the sense that, under some bijective mapping $\sigma_h : \mathcal{S}_h \to \mathcal{S}_h^*$,

$$\forall s^* \in \mathcal{S}_h^*, \quad \Pr_{x \sim \mathcal{Q}(s^*, P_h^e)}(\phi_h'(x) = \sigma_h^{-1}(\phi_h^*(x))) \geq 1 - \epsilon_0. \quad (17)$$

4 Simulation Results

Table 1: Results of toy environment simulation, with $H = 30, M = 128$, averaged over 20 random seeds. See text of Section 4, and Appendix D for further details.

Technique	Avg. Encoder Acc. ($\lvert\tau_A\rvert = \lvert\tau_B\rvert = 500$)	" " 1000	" " 5000
CRAFT	86.4%	**97.7%**	**>99.9%**
Single-obs. classification	67.8%	68.7%	69.7%
Paired-obs. classification	**87.4%**	86.1%	82.1%

We test CRAFT on a toy environment which captures CRAFT's ability to distinguish controllable features in the observation space from time-correlated uncontrollable features. In the environment, $s_1^* = 0, \mathcal{A} = \mathcal{S}_{h>1}^* = \{0, 1\}$ and $s_{h+1}^* = a_h$; in other words, the agent can simply set the next latent state using the action. The exogenous state consists of $M-1$ factors: $e = (e^1, e^2, ..., e^{M-1})$. Each exogenous factor is a two-state Markov Chain: for $e^2, ...e^{M-1}$, the initial state distribution and state transition probabilities are arbitrary parameters chosen uniformly at random for each chain, while e^1 has $\Pr(e_1^1 = 0) = 0.5$ and transition probabilities of zero. The observation $x_h \in \{0, 1\}^M$ consists of s_h^* concatenated with $(s_h^*$ XOR $e_h^i)$, for each $i \in [H-1]$. Additionally, at each timestep, the order of s_h^* and the other factors is *permuted* by some arbitrary permutation which depends on h. The hypothesis classes are $\Phi_h := \{(x_h) \to (x_h)_i | i \in [M]\}$. The representation learning problem is then to determine, for each h, which of the M components of the observation x_h is the controllable factor s_h^* (or, failing at that, to find a component corresponding to a $(s_h^*$ XOR $e_h^i)$ where e_h^i is low-entropy, so the encoder imperfect but still useful). Agent A selects actions uniformly at random, while for agent B, $\Pr(a_h = s_h^*) = 3/4$.

Results are shown in Table 1. The setting is designed to prevent various "shortcuts" to learning an encoder from working. Simply choosing the component of x_h that best predicts the policy ("Single-observation classification" in Table 1) will not work, because at any sufficiently large timestep h, the latent state distributions of the two policies are essentially identical (with a total-variation gap of 2^{-h}). Furthermore, given observations of a *pair* of sequential timesteps (x_h, x_{h+1}), choosing the components of x_h and x_{h+1}, respectively, that *together* best predict the agent also will not work ("Paired-observation classification"). In particular, the "distractor" features $(s_h^*$ XOR $e_h^1)$ and $(s_{h+1}$ XOR $e_{h+1}^1)$ are, taken together, about as informative about the *agent's identity* as s_h^* and s_{h+1}^*, but provide *no* information about the latent state s_h^* or s_{h+1}^*. In Table 1, we see that, given sufficient data ($\geq$ 1000 trajectories for each agent), CRAFT is capable of learning highly-accurate encoders in this setting, while these two "shortcut" techniques are not. In particular, while the "Paired-observation classification" shortcut is about as effective as CRAFT in the very-low data regime, its performance plateaus (and even seems to drop) as more data becomes available. (The *drop* in performance is likely because the adversarially-designed "distractor" features $(s_h^*$ XOR $e_h^1)$ and $(s_{h+1}^*$ XOR $e_{h+1}^1)$ are more likely to be chosen by this method as more data becomes available.)

5 Related Works

Action-free representation learning Many prior works have tackled action-free representation learning in practical scenarios, demonstrating empirically-validated methods. Common approaches utilize observation reconstruction losses (Seo et al., 2022) or temporal contrastive losses (Nair et al., 2023). Some of these works infer "latent actions" by finding a compact representation that is highly informative for predicting forward dynamics (Edwards et al., 2019; Menapace et al., 2021; Ye et al., 2023; Schmidt & Jiang, 2024) Another line of work augments large action-free offline datasets with significantly smaller action-labeled datasets. For example, an offline dataset action-free dataset can be used to train a goal-conditioned value function (Xu et al., 2022; Ma et al., 2023; Ghosh et al., 2023; Park et al., 2023). Alternatively, an inverse-dynamics model can be learned from the action-labelled data to "fill in" missing actions (Schmeckpeper et al., 2021; Zheng et al., 2023; Baker et al., 2022). By contrast, in this work we are interested in provable sample-efficiency of representation learning, and assume no access to action-labeled data during pretraining.

Learning in Ex-BMDPs. As discussed throughout this work, numerous prior works consider the Ex-BMDP model (Efroni et al., 2022; Mhammedi et al., 2024), including in the offline setting (Islam et al., 2023; Lamb et al., 2023; Levine et al., 2024). Misra et al. (2024) in particular demonstrates a hardness result: that Ex-BMDP latent representations cannot be learned in general from offline action-free data. In this work, we demonstrate a special case where this representation learning problem is in fact tractable: the case where offline data from multiple diverse agents are available.

6 Discussion and Limitations

One major assumption of this work (as well as Misra et al. (2024); Islam et al. (2023) and other prior works) is that offline data are collected by a policy which acts independently of observation noise. This assumption stems from the fact that, if noise features influence the behavioral policy, they (indirectly) influence the latent-state dynamics of the agent: these noise features may then be erroneously captured in the learned representation. However, in real-world settings, it may actually be beneficial to capture such features in the learned representation: if "expert" agents are relying on some uncontrollable feature, this feature may be relevant to the expert agents' reward functions, and may therefore also be relevant to downstream tasks for which our learned representations will be used. Therefore, the noise-independent policy assumption might not be necessary in practice.

An additional restrictive assumption of this work is that the latent dynamics are deterministic, and that each episode starts at the same latent state s_1^*. However, this assumption is also essentially present even in the best-known result for provably sample-efficient Ex-BMDP representation learning in the *online* episodic setting (Efroni et al., 2022) – that work does allow for *rare* departures from deterministic dynamics, however, and it may be possible to adapt the analysis of CRAFT to that setting as well, although we have focused on the strictly-deterministic case here for ease of presentation. Mhammedi et al. (2024) proposes an online algorithm for learning Ex-BMDPs with nondeterministic latent dynamics, but that work assumes "simulator access": the ability to reset the environment to any previously-visited observation. Several works (Lamb et al., 2023; Levine et al., 2024; Islam et al., 2023) consider learning Ex-BMDPs from offline data (with action labels) without assuming restarts to s_1^*: these works are "practical" algorithms that do not provide sample-complexity guarantees. A similar "practical" algorithm for the action-free, multiple-agent setting based on the ideas presented in this work may also be possible.

The assumption that two policies differ substantially at *every* latent state may also be impractical. One direction for future work may be to leverage data from *several* agents, such that it is more likely that *some* agent has a distinct behavior at each latent state.

While access to training oracles is a common assumption in representation learning (Agarwal et al., 2020; Efroni et al., 2022; Uehara et al., 2022), the optimization of Equation 13, on a *discretized* domain, may be troublesome in practice. Additionally, the sample complexity bounds in Equation 16, while polynomial, may not be optimal: these issues are potential directions for future work.

Acknowledgments

A portion of this research has taken place in the Learning Agents Research Group (LARG) at the Artificial Intelligence Laboratory, The University of Texas at Austin. LARG research is supported in part by the National Science Foundation (FAIN-2019844, NRT-2125858), the Office of Naval Research (N00014-18-2243), Army Research Office (W911NF-23-2-0004, W911NF-17-2-0181), DARPA (Cooperative Agreement HR00112520004 on Ad Hoc Teamwork), Lockheed Martin, and Good Systems, a research grand challenge at the University of Texas at Austin. The views and conclusions contained in this document are those of the authors alone. 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 (FAIN-2019844). Amy Zhang and Alexander Levine are supported by National Science Foundation (2340651) and Army Research Office (W911NF-24-1-0193).

References

Alekh Agarwal, Sham Kakade, Akshay Krishnamurthy, and Wen Sun. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Bowen Baker, Ilge Akkaya, Peter Zhokov, Joost Huizinga, Jie Tang, Adrien Ecoffet, Brandon Houghton, Raul Sampedro, and Jeff Clune. Video pretraining (vpt): Learning to act by watching unlabeled online videos. *Advances in Neural Information Processing Systems*, 35:24639–24654, 2022.

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.

Ashley Edwards, Himanshu Sahni, Yannick Schroecker, and Charles Isbell. Imitating latent policies from observation. In *International conference on machine learning*, pp. 1755–1763. PMLR, 2019.

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*, 2022. URL https://openreview.net/forum?id=RQLLzMCefQu.

Dibya Ghosh, Chethan Anand Bhateja, and Sergey Levine. Reinforcement learning from passive data via latent intentions. In *International Conference on Machine Learning*, pp. 11321–11339. PMLR, 2023.

Riashat Islam, Manan Tomar, Alex Lamb, Yonathan Efroni, Hongyu Zang, Aniket Rajiv Didolkar, Dipendra Misra, Xin Li, Harm Van Seijen, Remi Tachet Des Combes, et al. Principled offline rl in the presence of rich exogenous information. In *International Conference on Machine Learning*, pp. 14390–14421. PMLR, 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*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=TNocbXm5MZ.

Alexander Levine, Peter Stone, and Amy Zhang. Multistep inverse is not all you need. *Reinforcement Learning Journal*, 2:884–925, 2024.

Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=YJ7o2wetJ2.

Willi Menapace, Stephane Lathuiliere, Sergey Tulyakov, Aliaksandr Siarohin, and Elisa Ricci. Playable video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10061–10070, 2021.

Zakaria Mhammedi, Dylan J Foster, and Alexander Rakhlin. The power of resets in online reinforcement learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=7sACcaOmGi.

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.

Dipendra Misra, Akanksha Saran, Tengyang Xie, Alex Lamb, and John Langford. Towards principled representation learning from videos for reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3mnWvUZIXt.

Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning*, pp. 892–909. PMLR, 2023.

Seohong Park, Dibya Ghosh, Benjamin Eysenbach, and Sergey Levine. HIQL: Offline goal-conditioned RL with latent states as actions. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=cLQCCtVDuW.

Karl Schmeckpeper, Oleh Rybkin, Kostas Daniilidis, Sergey Levine, and Chelsea Finn. Reinforcement learning with videos: Combining offline observations with interaction. In Jens Kober, Fabio Ramos, and Claire Tomlin (eds.), *Proceedings of the 2020 Conference on Robot Learning*, volume 155 of *Proceedings of Machine Learning Research*, pp. 339–354. PMLR, 16–18 Nov 2021. URL https://proceedings.mlr.press/v155/schmeckpeper21a.html.

Dominik Schmidt and Minqi Jiang. Learning to act without actions. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=rvUq3cxpDF.

Younggyo Seo, Kimin Lee, Stephen L James, and Pieter Abbeel. Reinforcement learning with action-free pre-training from videos. In *International Conference on Machine Learning*, pp. 19561–19579. PMLR, 2022.

Masatoshi Uehara, Xuezhou Zhang, and Wen Sun. Representation learning for online and offline RL in low-rank MDPs. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=J4iSIR9fhY0.

Haoran Xu, Li Jiang, Jianxiong Li, and Xianyuan Zhan. A policy-guided imitation approach for offline reinforcement learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=CKbqDtZnSc.

Weirui Ye, Yunsheng Zhang, Pieter Abbeel, and Yang Gao. Become a proficient player with limited data through watching pure videos. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Sy-o2N0hF4f.

Qinqing Zheng, Mikael Henaff, Brandon Amos, and Aditya Grover. Semi-supervised offline reinforcement learning with action-free trajectories. In *International conference on machine learning*, pp. 42339–42362. PMLR, 2023.

Supplementary Materials

The following content was not necessarily subject to peer review.

A Hypothesis Classes and Realizability Assumptions

As mentioned in Section 2.5, we assume access to hypothesis classes of encoder functions $\Phi_{1:H}$. We make a realizability assumption: that is, the true encoder $\phi_h^* \in \Phi_h$. Moreover, we assume that for any arbitrary permutation σ, $\sigma(\phi_h^*(x)) \in \Phi_h$ – this allows us to train an encoder in Φ_h on datasets of observations representing each latent state without knowing the "correct" ordering of the latent states. Similar realizability assumptions are common in representation learning literature for structured MDPs (Du et al., 2019; Efroni et al., 2022; Misra et al., 2020; 2024; Uehara et al., 2022; Agarwal et al., 2020).

We use a second set of hypothesis classes, $\mathcal{G}_h \subseteq \mathcal{X}_h \to \{0, 1\}$, for which, for any *pair* of latent states $s_h^*, s_h'^*$, there exists some $g \in \mathcal{G}_h$ that can perfectly distinguish observations of s_h^* from observations of $s_h'^*$. In our sample-complexity results, we assume that $|\mathcal{G}_h| \leq |\Phi_h|$. In this work, we are chiefly concerned with sample-complexity: we make use of training oracles for a variety of loss functions which may not be tractable to optimize in practice. See Section 6 for further discussion.

B Algorithm

The full CRAFT algorithm is presented as Algorithm 2.

C Proofs

In this section, we prove the correctness and sample complexity bounds of CRAFT presented in Theorem 3.1. First, though, we prove various lemmas the will be helpful in proving the final result.

C.1 Preliminary Note

Recall Equation 1 in the main text:

$$\begin{aligned}\Pr(\tau_A, \tau_B) = &\Pr(\phi^*(\tau_A), \phi^*(\tau_B)) \cdot \Pr_{P_1^e, \mathcal{T}^e}(\phi^e(\tau_A)) \cdot \Pr_{P_1^e, \mathcal{T}^e}(\phi^e(\tau_B)) \\ &\cdot \Pr_Q(\tau_A | \phi^*(\tau_A), \phi^e(\tau_A)) \cdot \Pr_Q(\tau_B | \phi^*(\tau_B), \phi^e(\tau_B))\end{aligned}$$

Throughout our proofs, we will make use of this assumption in the following way: we will treat the controllable latent state trajectories $\phi^*(\tau_A), \phi^*(\tau_B)$ as *fixed but arbitrary*, not as random variables, and treat the exogenous noise Markov chains $\phi^e(\tau_A), \phi^e(\tau_B)$ and the emission function $\mathcal{Q}$ as the only random variables. Then, if the algorithm succeeds with high probability for any such *fixed, arbitrary* $\phi^*(\tau_A), \phi^*(\tau_B)$, we can conclude by the independence assumption that it also succeeds with high probability under any data-generating process for which Equation 1 holds.

C.2 Concentration Lemmas

In this section, we present concentration bounds on the loss functions used in Algorithm 2. We start with the log-odds loss given in Equation 13:

Lemma C.1. . *Given m distributions $\mathcal{D}_1, ..., \mathcal{D}_m \in \mathcal{P}(\mathcal{X})$, each with two corresponding positive integers a_i, b_i, for $i \in [m]$, let $A_i \sim \mathcal{D}^{a_i}$ and $B_i \sim \mathcal{D}^{b_i}$ be two multi-sets consisting of a_i and b_i i.i.d. samples from $\mathcal{D}_i$, respectively. Then, for any $\xi > 0$ and $n_\Xi \in \mathbb{N}_+$ such that $\forall i$, $|\ln(a_i/b_i)| \leq \frac{n_\Xi \xi}{2}$, let $\Xi = \{-\frac{n_\Xi \xi}{2}, -\frac{n_\Xi \xi}{2} + \xi, -\frac{n_\Xi \xi}{2} + 2\xi, ..., \frac{n_\Xi \xi}{2}\}$. Further, let $\bar{c}_i \in \Xi$ be the smallest value in*

Algorithm 2 CRAFT

Require: Trajectory datasets τ_A, τ_B, known lower-bounds α, η, and ν, encoder function classes $\Phi_h \subseteq \mathcal{X}_h \to N_s$ and classification function class $\mathcal{G}_h \subseteq \mathcal{X}_h \to \{0,1\}$.

$\alpha \leftarrow \min(1, \alpha)$.
Let $\xi := \alpha/4$; $n_\Xi := \lceil 8\ln(\eta^{-1}-1)/\alpha \rceil$.
$\eta \leftarrow 1/(1+e^{n_\Xi \alpha/8})$.
Let $\Xi := \{-\frac{n_\Xi \xi}{2}, -\frac{n_\Xi \xi}{2}+\xi, -\frac{n_\Xi \xi}{2}+2\xi, ..., \frac{n_\Xi \xi}{2}\}$.
Initialize $\mathcal{S}_1 := \{s_1\}$, $D_{A,1}(s_1) := [|\tau_A|]$, $D_{B,1}(s_1) := [|\tau_B|]$. `\\ First timestep should have a single state rep., associated with every trajectory index.`
Let $\phi'_1 := \mathcal{X}_1 \to 0$.
for $h \in \{1, 2, ..., H-1\}$ **do**
  Initialize $\mathcal{S}_{h+1} := \{\}$.
  Let the inverse-actor-prediction function class $\mathcal{F}_h \subseteq \mathcal{X} \times \mathcal{X} \to \Xi$ be composed as $\mathcal{F}_h = \Phi_h \times \Phi_{h+1} \times (N_s^2 \to \Xi)$.
  Find the $f_h \in \mathcal{F}_h$ which minimizes Equation 13.
  Let $q_{thresh.} := \frac{h\nu}{8H}$.
  for $s_{pred} \in \mathcal{S}_h$ **do**
    Initialize merged_already$(s) :=$ False for all $s \in \mathcal{S}_{h+1}$.
    Initialize $\mathcal{S}_{new} := \{\}$
    Let pairs$(s_{pred}) := (\tau_A)_{h:\,h+1}[D_{A,h}(s_{pred})] \cup (\tau_B)_{h:\,h+1}[D_{B,h}(s_{pred})]$. `\\ All transitions which start at` s_{prev}.
    $\forall j \in \{0, ..., n_\Xi\}$, Let pred_succ$[j] := \{(x_h, x_{h+1}) \in$ pairs$(s_{pred}) | f(x_h, x_{h+1}) = j\xi - \frac{n_\Xi \xi}{2}\}$.
    Initialize $j := 0$
    while $j \leq n_\Xi$ **do**
      if |pred_succ$[j]| \geq q_{thresh.}(|\tau_A| + |\tau_B|)$ **then**
        Let j' be the minimum integer $> j$ such that |pred_succ$[j']| < q_{thresh.}(|\tau_A| + |\tau_B|)$, or n_Ξ if no such integer exists.
        Let $\mathcal{D}_{new\_pairs} := \{x' | (x,x') \in \biguplus_{k=\max(0,j-1)}^{j'}$ pred_succ$[k]\}$, $\mathcal{D}_{new} := \{x' | (x,x') \in \mathcal{D}_{new\_pairs}\}$
        Initialize new_state? $\leftarrow$ True.
        for $s \in \mathcal{S}_{h+1}$, such that merged_already$(s) ==$ False **do**
          Let $\mathcal{D}_s := (\tau_A)_{h+1}[D_{A,h+1}(s)] \uplus (\tau_B)_{h+1}[D_{B,h+1}(s)]$
          Train a classifier $g \in \mathcal{G}$ to distinguish $\mathcal{D}_{new}$ and $\mathcal{D}_s$, with loss $\mathcal{L}(g)$ given in Equation 53.
          if the loss $\mathcal{L}(g)$ on $\mathcal{D}_{new}$ and $\mathcal{D}_s$ is > 0.5 **then**
            Append to $D_{A,h+1}(s)$ indices of trajectories in τ_A that observations in $\mathcal{D}_{new}$ are from.
            Append to $D_{B,h+1}(s)$ indices of trajectories in τ_B that observations in $\mathcal{D}_{new}$ are from.
            merged_already?$(s) \leftarrow$ True
            new_state? $\leftarrow$ False
            break.
          end if
        end for
        if new_state? **then**
          Add new state s_{new} to $\mathcal{S}_{new}$
          Initialize $D_{A,h+1}(s_{new})$ as indices of trajectories in τ_A that observations in $\mathcal{D}_{new}$ are from.
          Initialize $D_{B,h+1}(s_{new})$ as indices of trajectories in τ_B that observations in $\mathcal{D}_{new}$ are from.
        end if
        $j \leftarrow j' + 2$
      else
        $j \leftarrow j + 1$
      end if
    end while
    $\mathcal{S}_{h+1} := \mathcal{S}_{h+1} \cup \mathcal{S}_{new}$
  end for
  $\phi'_{h+1} := \arg\min_{\phi \in \Phi_{h+1}} \sum_{s \in \mathcal{S}_{h+1}} \left[\frac{1}{|\mathcal{D}_s|} \sum_{x \in \mathcal{D}_s} (1 - \mathbb{1}_{(\phi(x)=s)}) \right]$, where $\mathcal{D}_s := (\tau_A)_{h+1}[D_{A,h+1}(s)] \uplus (\tau_B)_{h+1}[D_{B,h+1}(s)]$
end for
Return: $\phi'_1, ...\phi'_H$

Ξ greater than or equal to $\ln(a_i/b_i)$, and $\underline{c}_i \in \Xi$ be the largest value in Ξ less than or equal to $\ln(a_i/b_i)$. Also, assume that $\forall i, \frac{a_i+b_i}{\sum_{i'=1}^m a_{i'}+b_{i'}} \geq \nu$.

Given any function $f \in \mathcal{X} \to \Xi$, define:

$$\mathcal{L}(f) := \sum_{i=1}^{m} \left[\sum_{x \in A_i} \ln(e^{-f(x)} + 1) + \sum_{x \in B_i} \ln(e^{f(x)} + 1) \right]. \tag{18}$$

Further, define:

$$\mathcal{L}_{ref} := \sum_{i=1}^{m} \left[\min_{c_i \in \{\bar{c}_i, \underline{c}_i\}} a_i \ln\left(e^{-c_i} + 1\right) + b_i \ln\left(e^{c_i} + 1\right) \right], \tag{19}$$

and let $\eta := (e^{n\Xi\xi/2} + 1)^{-1}$. For any ϵ and δ, if:

$$\forall i \in [m], \;\; a_i + b_i \geq \frac{50 \ln(2/\delta) \ln^2(1/\eta)}{\nu \epsilon^2 \eta^2 \xi^4} \tag{20}$$

*then the probability that **both** $\mathcal{L} \leq \mathcal{L}_{ref}$, **and**:*

$$\exists i \in [m] : \;\; \left| \{x \in A_i \uplus B_i | f(x) \notin \{\underline{c}_i, \bar{c}_i\} \right| > \epsilon(a_i + b_i) \tag{21}$$

is at most δ.

Proof. We can define

$$\mathcal{L}^i_{\text{ref}} := \min_{c_i \in \{\bar{c}_i, \underline{c}_i\}} a_i \ln\left(e^{-c_i} + 1\right) + b_i \ln\left(e^{c_i} + 1\right), \tag{22}$$

So that $\mathcal{L}_{\text{ref}} = \sum_{i=1}^m \mathcal{L}^i_{\text{ref}}$, and also define

$$\mathcal{L}^i_{pop}(f) := a_i \left(\mathop{\mathbb{E}}_{x \sim \mathcal{D}_i} \ln(e^{-f(x)} + 1) \right) + b_i \left(\mathop{\mathbb{E}}_{x \sim \mathcal{D}_i} \ln(e^{f(x)} + 1) \right) \tag{23}$$

and $\mathcal{L}_{\text{pop}} := \sum_{i=1}^m \mathcal{L}^i_{\text{pop}}$.

First, we consider the "population loss" for each distribution:

$$\begin{aligned}
\mathcal{L}^i_{pop}(f) &= a_i \left(\mathop{\mathbb{E}}_{x \sim \mathcal{D}_i} \ln(e^{-f(x)} + 1) \right) + b_i \left(\mathop{\mathbb{E}}_{x \sim \mathcal{D}_i} \ln(e^{f(x)} + 1) \right) \\
&= a_i \left(\sum_{\zeta \in \Xi} \Pr_{x \sim \mathcal{D}_i}(f(x) = \zeta) \cdot \ln(e^{-\zeta} + 1) \right) \\
&\quad + b_i \left(\sum_{\zeta \in \Xi} \Pr_{x \sim \mathcal{D}_i}(f(x) = \zeta) \cdot \ln(e^{\zeta} + 1) \right) \\
&= \sum_{\zeta \in \Xi} \Pr_{x \sim \mathcal{D}_i}(f(x) = \zeta) \cdot \left(a_i \ln(e^{-\zeta} + 1) + b_i \ln(e^{\zeta} + 1) \right) \\
&= \sum_{\zeta \in \Xi} \Pr_{x \sim \mathcal{D}_i}(f(x) = \zeta) \cdot a_i \left(\left(1 + \frac{b_i}{a_i}\right) \ln(e^{\zeta} + 1) - \zeta \right)
\end{aligned} \tag{24}$$

We can define $h_\gamma(\zeta) := \left(1 + \gamma^{-1}\right) \ln(e^\zeta + 1) - \zeta$, so that

$$\mathcal{L}^i_{pop}(f) = a_i \sum_{\zeta \in \Xi} \Pr_{x \sim \mathcal{D}_i}(f(x) = \zeta) \cdot h_{a_i/b_i}(\zeta) \tag{25}$$

Now, note that:

$$h'_\gamma(\zeta) = \left(1+\gamma^{-1}\right)\frac{e^\zeta}{e^\zeta+1} - 1 \tag{26}$$

and

$$h''_\gamma(\zeta) = \left(1+\gamma^{-1}\right)\frac{e^\zeta}{(e^\zeta+1)^2}. \tag{27}$$

Then, we see that $h_\gamma(\zeta)$ is a convex function, with a global minimum at $\zeta = \ln(\gamma)$, and second-derivative at least

$$\left(1+\gamma^{-1}\right)\frac{e^{n_\Xi \xi/2}}{(e^{n_\Xi\xi/2}+1)^2} \quad \left(= \left(1+\gamma^{-1}\right)\frac{e^{-n_\Xi \xi/2}}{(e^{-n_\Xi\xi/2}+1)^2}\right) \tag{28}$$

everywhere on the interval $[-n_\Xi\xi/2, n_\Xi\xi/2]$.

Due to the convexity of $h_{a_i/b_i}(\zeta)$, we have, $\forall j > 0$,

$$a_i{\cdot}h_{a_i/b_i}(\ln(a_i/b_i)) \leq \min(a_i{\cdot}h_{a_i/b_i}(\bar{c}_i), a_i{\cdot}h_{a_i/b_i}(\underline{c}_i)) \,(= \mathcal{L}^i_{ref}) \leq a_i{\cdot}h_{a_i/b_i}(\bar{c}_i) < a_i{\cdot}h_{a_i/b_i}(\bar{c}_i{+}j\xi)$$

and, similarly, $\forall j > 0$:

$$a_i{\cdot}h_{a_i/b_i}(\ln(a_i/b_i)) \leq \min(a_i{\cdot}h_{a_i/b_i}(\bar{c}_i), a_i{\cdot}h_{a_i/b_i}(\underline{c}_i)) \,(= \mathcal{L}^i_{ref}) \leq a_i{\cdot}h_{a_i/b_i}(\underline{c}_i) < a_i{\cdot}h_{a_i/b_i}(\underline{c}_i{-}j\xi).$$

In particular, by a second-order Taylor bound, we have that, for $j > 0$:

$$\begin{aligned}\mathcal{L}^i_{ref} &\leq a_i \cdot h_{a_i/b_i}(\bar{c}_i + j\xi) - a_i \cdot \left(1+(a_i/b_i)^{-1}\right)\frac{e^{n_\Xi\xi/2}}{(e^{n_\Xi\xi/2}+1)^2}\cdot\frac{(j\xi)^2}{2} \\ &\leq a_i \cdot h_{a_i/b_i}(\bar{c}_i + j\xi) - (a_i+b_i)\frac{e^{n_\Xi\xi/2}}{(e^{n_\Xi\xi/2}+1)^2}\cdot\frac{\xi^2}{2}\end{aligned} \tag{29}$$

and similarly for $a_i \cdot h_{a_i/b_i}(\underline{c}_i - j\xi)$:

$$\mathcal{L}^i_{ref} \leq a_i \cdot h_{a_i/b_i}(\underline{c}_i - j\xi) - (a_i+b_i)\frac{e^{n_\Xi\xi/2}}{(e^{n_\Xi\xi/2}+1)^2}\cdot\frac{\xi^2}{2}. \tag{30}$$

In particular, by Equation 25,

$$\mathcal{L}^i_{pop}(f) \geq \mathcal{L}^i_{ref} + \Pr_{x\sim\mathcal{D}_i}(f(x) \notin \{\underline{c}_i, \bar{c}_i\}) \cdot (a_i+b_i)\frac{e^{n_\Xi\xi/2}}{(e^{n_\Xi\xi/2}+1)^2}\cdot\frac{\xi^2}{2}. \tag{31}$$

In terms of η, this is:

$$\begin{aligned}\mathcal{L}^i_{pop}(f) &\geq \mathcal{L}^i_{ref} + \Pr_{x\sim\mathcal{D}_i}(f(x) \notin \{\underline{c}_i, \bar{c}_i\}) \cdot (a_i+b_i)\,(\eta-\eta^2)\cdot\frac{\xi^2}{2} \\ &\geq \mathcal{L}^i_{ref} + \Pr_{x\sim\mathcal{D}_i}(f(x) \notin \{\underline{c}_i, \bar{c}_i\}) \cdot (a_i+b_i)\cdot\frac{\eta\xi^2}{4}\end{aligned} \tag{32}$$

where we use the fact that $\eta \leq 1/2$ in the last inequality. This gives us:

$$\Pr_{x\sim\mathcal{D}_i}(f(x) \notin \{\underline{c}_i, \bar{c}_i\}) \leq \frac{4(\mathcal{L}^i_{pop}(f) - \mathcal{L}^i_{ref})}{(a_i+b_i)\eta\xi^2} \tag{33}$$

Because $\mathcal{L}^i_{pop} - \mathcal{L}^i_{ref} \geq 0$, this implies:

$$\begin{aligned}&\forall i \in [m], \\ &(a_i+b_i)\cdot\Pr_{x\sim\mathcal{D}_i}(f(x) \notin \{\underline{c}_i, \bar{c}_i\}) \leq \frac{4(\mathcal{L}^i_{pop}(f) - \mathcal{L}^i_{ref})}{\eta\xi^2} \leq \frac{4\sum_{i=1}^m(\mathcal{L}^i_{pop}(f) - \mathcal{L}^i_{ref})}{\eta\xi^2} \\ &\qquad = \frac{4(\mathcal{L}_{pop}(f) - \mathcal{L}_{ref})}{\eta\xi^2}\end{aligned} \tag{34}$$

Meanwhile, from Equations 18 and 23 applying (one-sided) Hoeffding's lemma gives us, with probability at least $1-\delta/2$:

$$\mathcal{L}(f) - \mathcal{L}_{pop}(f) + \sqrt{\sum_{i=1}^{m}(a_i+b_i)}\ln(1/\eta)\sqrt{2\ln(2/\delta)} \geq 0. \tag{35}$$

which implies, by assumption:

$$\forall i,\ \mathcal{L}(f) - \mathcal{L}_{pop}(f) + \sqrt{a_i+b_i}\sqrt{1/\nu}\ln(1/\eta)\sqrt{2\ln(2/\delta)} \geq 0. \tag{36}$$

Combining Equations 34 and 36 gives, with probability at least $1-\delta/2$, we have that $\forall i \in [m]$,

$$(a_i+b_i)\Pr_{x\sim\mathcal{D}_i}(f(x) \notin \{\underline{c}_i, \bar{c}_i\}) \leq \frac{4(\mathcal{L}(f)-\mathcal{L}_{ref})}{\eta\xi^2} + \frac{4(\sqrt{a_i+b_i})\ln(1/\eta)\sqrt{2\ln(2/\delta)}}{\sqrt{\nu}\eta\xi^2}. \tag{37}$$

Then, with probability at least $1-\delta/2$, the condition $\mathcal{L}(f) \leq \mathcal{L}_{ref}$ implies:

$$\forall i \in [m],\ \ (a_i+b_i)\Pr_{x\sim\mathcal{D}_i}(f(x) \notin \{\underline{c}_i, \bar{c}_i\}) \leq \frac{4\sqrt{a_i+b_i}\ln(1/\eta)\sqrt{2\ln(2/\delta)}}{\sqrt{\nu}\eta\xi^2}. \tag{38}$$

We can apply Hoeffding's lemma *once for each* $i \in [m]$, to the binary variable of whether on not $f(x) \in \{\underline{c}_i, \bar{c}_i\}$, where x is sampled (a_i+b_i) times to produce the dataset $A_i \cup B_i$. By union bound, we have, with probability at least $1-\delta$, $\mathcal{L}(f) \leq \mathcal{L}_{ref}$ implies:

$$\begin{aligned}
\forall i \in [m],\ \ \left|\{x \in A_i \cup B_i | f(x) \notin \{\underline{c}_i, \bar{c}_i\}\right| \leq \\
\sqrt{(a_i+b_i)\ln(2m/\delta)/2} + \frac{4\sqrt{a_i+b_i}\ln(1/\eta)\sqrt{2\ln(2/\delta)}}{\sqrt{\nu}\eta\xi^2} \leq \\
\sqrt{(a_i+b_i)}\sqrt{2}\left(\frac{\sqrt{\ln(2m/\delta)}}{2} + \frac{4\sqrt{\ln(2/\delta)}\ln(1/\eta)}{\sqrt{\nu}\eta\xi^2}\right) \leq \\
\sqrt{(a_i+b_i)}\sqrt{2}\frac{\ln(1/\eta)}{\eta\xi^2}\left(\frac{\sqrt{\ln(2/\delta)}}{2} + \frac{\sqrt{\ln(m)}}{2} + \frac{4\sqrt{\ln(2/\delta)}}{\sqrt{\nu}}\right),
\end{aligned} \tag{39}$$

where in the last line, we used triangle inequality and the fact that $\eta = 1/(e^{n\Xi\xi/2}+1) \leq 1/(e^{\xi/2}+1)$, which in turn implies:

$$\frac{\ln(1/\eta)}{\eta\xi^2} \geq \frac{(e^{\xi/2}+1)\ln(e^{\xi/2}+1)}{\xi^2} > 1 \quad (\forall \xi > 0). \tag{40}$$

Note that, because each a_i+b_i contains at least a ν-fraction of the total $\sum_i^m a_i+b_i$, we must have $m \leq 1/\nu$. Then:

$$\frac{\sqrt{\ln(m)}}{2} \leq \frac{\sqrt{\ln(1/\nu)}}{2} \leq \frac{1}{2\sqrt{e}\sqrt{\nu}} \leq \frac{1}{2\sqrt{e}\sqrt{\ln(2)}}\frac{\sqrt{\ln(2/\delta)}}{\sqrt{\nu}} \leq \frac{\sqrt{\ln(2/\delta)}}{2\sqrt{\nu}} \tag{41}$$

Therefore (and noting $\nu < 1$), we can combine terms in Equation 39 to conclude:

$$\forall i \in [m],\ \ \left|\{x \in A_i \cup B_i | f(x) \notin \{\underline{c}_i, \bar{c}_i\}\right| \leq \frac{5\sqrt{2(a_i+b_i)\ln(2/\delta)}\ln(1/\eta)}{\sqrt{\nu}\eta\xi^2}. \tag{42}$$

Now, to ensure $\forall i \in [m],\ \ \left|\{x \in A_i \cup B_i | f(x) \notin \{\underline{c}_i, \bar{c}_i\}\right| \leq \epsilon(a_i+b_i)$, we need,

$$\forall i \in [m],\ \ \frac{5\sqrt{2(a_i+b_i)\ln(2/\delta)}\ln(1/\eta)}{\sqrt{\nu}\eta\xi^2} \leq \epsilon(a_i+b_i) \tag{43}$$

or:

$$\forall i \in [m], \quad \frac{50 \ln(2/\delta) \ln^2(1/\eta)}{\nu \epsilon^2 \eta^2 \xi^4} \leq a_i + b_i \tag{44}$$

as provided by Equation 20. Note that because the implication

$$\mathcal{L}(f) \leq \mathcal{L}_{ref} \rightarrow \forall i \in [m], \quad \left| \{x \in A_i \cup B_i | f(x) \notin \{\underline{c}_i, \bar{c}_i\} \right| \leq \epsilon(a_i + b_i) \tag{45}$$

holds with probability at least $(1 - \delta)$, this implication can only be *broken*, by the case that

$$\mathcal{L}(f) \leq \mathcal{L}_{ref} \wedge \exists i \in [m] : \left| \{x \in A_i \cup B_i | f(x) \notin \{\underline{c}_i, \bar{c}_i\} \right| > \epsilon(a_i + b_i) \tag{46}$$

with probability at most δ. □

Corollary C.2. *Let $\mathcal{D}^*(s^*_h, s^*_{h+1})$ be the multiset of observation pairs (x_h, x_{h+1}) from both τ_A and τ_B in Algorithm 1, such that $\phi^*_h(x_h) = s^*_h$ and $\phi^*_{h+1}(x_{h+1}) = s^*_{h+1}$, and let $\mathcal{D}^*_A(s^*_h, s^*_{h+1})$ and $\mathcal{D}^*_B(s^*_h, s^*_{h+1})$ be the elements of $\mathcal{D}^*(s^*_h, s^*_{h+1})$ originating from τ_A and τ_B respectively. Further, let $\bar{c}_{s^*_h, s^*_{h+1}} \in \Xi$ be the smallest value in Ξ greater than or equal to $\ln(|\mathcal{D}^*_A(s^*_h, s^*_{h+1})|/|\mathcal{D}^*_B(s^*_h, s^*_{h+1})|)$, and $\underline{c}_{s^*_h, s^*_{h+1}} \in \Xi$ be the largest value in Ξ less than or equal to $\ln(|\mathcal{D}^*_A(s^*_h, s^*_{h+1})|/|\mathcal{D}^*_B(s^*_h, s^*_{h+1})|)$.*

Further, assume the realizability condition that $\phi^*_h \in \Phi_h$ and $\phi^*_{h+1} \in \Phi_{h+1}$.

If $\forall s^*_h, s^*_{h+1}$, such that s^*_h can transition to s^*_{h+1},

$$|\mathcal{D}^*(s^*_h, s^*_{h+1})| \geq \frac{50(\ln(2|\Phi|^2/\delta) + N_s^2 \ln(n_\Xi + 1)) \ln^2(1/\eta)}{\nu \epsilon^2 \eta^2 \xi^4} \tag{47}$$

then with probability at least $1 - \delta$, the function $f(\cdot)$ found in Line 10 of Algorithm 2 will be such that, $\forall s^*_h, s^*_{h+1}$, such that s^*_h can transition to s^*_{h+1},

$$\left| \{x \in \mathcal{D}^*(s^*_h, s^*_{h+1}) | f(x) \notin \{\underline{c}_{s^*_h, s^*_{h+1}}, \bar{c}_{s^*_h, s^*_{h+1}}\} \right| \leq \epsilon \left| \mathcal{D}^*(s^*_h, s^*_{h+1}) \right|. \tag{48}$$

Proof. By application of Lemma C.1 with $\delta' := \delta / (|\Phi|^2 \cdot (n_\Xi + 1)^{(N_s^2)}) \leq \delta / |\mathcal{F}_h|$, we have that, for any fixed hypothesis f', the probability that $\mathcal{L}(f') \leq \mathcal{L}_{ref}$ and Equation 48 is violated is at most $\delta / |\mathcal{F}_h|$. Then by union bound, the probability that any such f' exists in $\mathcal{F}_h$ is at most $1 - \delta$. However, by the realizability assumption, we know that an f^* exists in $\mathcal{F}$ which achieves loss $\mathcal{L}(f^*) = \mathcal{L}_{ref}$ and also that respects Equation 48. (In particular, this f^* is simply (ϕ^*_h, ϕ^*_{h+1}) composed with a mapping from the representations corresponding to each (s^*_h, s^*_{h+1}) to the corresponding $\underline{c}_{s^*_h, s^*_{h+1}}$ or $\bar{c}_{s^*_h, s^*_{h+1}}$ which minimizes Equation 22.) Therefore with probability at least $1 - \delta$, the $f \in \mathcal{F}_h$ which minimizes $\mathcal{L}(f)$ must respect Equation 48. □

We now give two simple results for classification under corrupted data. First though, we prove a minor claim, which is simply some "deferred algebra" for the lemmas which follow:

Proposition C.3. *Consider a multiset $\mathcal{Z} = \{z_1, ..., z_m\}$ of items $z_i \in [0, 1]$, and a modified multiset $\mathcal{Z}'$, also consisting of items in $[0, 1]$, such that the symmetric difference between $\mathcal{Z}$ and $\mathcal{Z}'$ has size at most k (that is, $\mathcal{Z}'$ can be constructed from $\mathcal{Z}$ by inserting and/or removing a total of at most k items). Then*

$$\left| \sum_{\mathcal{Z}} \frac{z}{|\mathcal{Z}|} - \sum_{\mathcal{Z}'} \frac{z'}{|\mathcal{Z}'|} \right| \leq \frac{k}{m} \tag{49}$$

Proof. Define $\mathcal{Z}_{removed}, \mathcal{Z}_{added}$, and $\mathcal{Z}_{kept}$ such that $\mathcal{Z} = \mathcal{Z}_{kept} + \mathcal{Z}_{removed}$, and $\mathcal{Z}' = \mathcal{Z}_{kept} + \mathcal{Z}_{added}$. Note that $k = |\mathcal{Z}_{removed}| + |\mathcal{Z}_{added}|$ and $m = |\mathcal{Z}_{removed}| + |\mathcal{Z}_{kept}|$. We first assume that $|\mathcal{Z}| \geq |\mathcal{Z}'|$. In other words, we assume $|\mathcal{Z}_{removed}| \geq |\mathcal{Z}_{added}|$. Then we can construct $\mathcal{Z}'$ from $\mathcal{Z}$ by (1) removing some arbitrary subset $\mathcal{Z}'_{removed} \subseteq \mathcal{Z}_{removed}$ from $\mathcal{Z}$, such that $|\mathcal{Z}'_{removed}| = |\mathcal{Z}_{added}|$; then (2) inserting the samples $\mathcal{Z}_{added}$; and then finally (3) removing the multiset $\mathcal{Z}''_{removed} = \mathcal{Z}_{removed} \setminus \mathcal{Z}'_{removed}$. Let the intermediate set constructed after step (2) be $\mathcal{Z}'' := (\mathcal{Z} \setminus \mathcal{Z}'_{removed}) \uplus \mathcal{Z}_{added} = \mathcal{Z}' \uplus \mathcal{Z}''_{removed}$. Note that $|\mathcal{Z}| = |\mathcal{Z}''|$, and

$$\begin{aligned} &\left|\sum_{\mathcal{Z}} \frac{z}{|\mathcal{Z}|} - \sum_{\mathcal{Z}''} \frac{z''}{|\mathcal{Z}''|}\right| = \left|\sum_{\mathcal{Z}} \frac{z}{|\mathcal{Z}|} - \sum_{\mathcal{Z}''} \frac{z''}{|\mathcal{Z}|}\right| = \\ &\frac{1}{|\mathcal{Z}|}\left|\sum_{\mathcal{Z}} z - \sum_{\mathcal{Z}''} z''\right| = \frac{1}{|\mathcal{Z}|}\left|\sum_{\mathcal{Z}'_{removed}} z - \sum_{\mathcal{Z}_{added}} z\right| \leq \frac{|\mathcal{Z}_{added}|}{|\mathcal{Z}|} \end{aligned} \tag{50}$$

Additionally, note that:

$$\begin{aligned} &\left|\sum_{\mathcal{Z}''} \frac{z''}{|\mathcal{Z}''|} - \sum_{\mathcal{Z}'} \frac{z'}{|\mathcal{Z}'|}\right| = \frac{1}{|\mathcal{Z}''|}\left|\sum_{\mathcal{Z}''} z'' - \frac{|\mathcal{Z}''|\sum_{\mathcal{Z}'} z'}{|\mathcal{Z}'|}\right| = \\ &\frac{1}{|\mathcal{Z}|}\left|\frac{|\mathcal{Z}'|\sum_{\mathcal{Z}'} z'}{|\mathcal{Z}'|} + \frac{|\mathcal{Z}''_{removed}|\sum_{\mathcal{Z}''_{removed}} z_r}{|\mathcal{Z}''_{removed}|} - \frac{|\mathcal{Z}''|\sum_{\mathcal{Z}'} z'}{|\mathcal{Z}'|}\right| = \\ &\frac{1}{|\mathcal{Z}|}\left|\frac{|\mathcal{Z}''_{removed}|\sum_{\mathcal{Z}''_{removed}} z_r}{|\mathcal{Z}''_{removed}|} - \frac{|\mathcal{Z}''_{removed}|\sum_{\mathcal{Z}'} z'}{|\mathcal{Z}'|}\right| = \\ &\frac{|\mathcal{Z}''_{removed}|}{|\mathcal{Z}|}\left|\frac{\sum_{\mathcal{Z}''_{removed}} z_r}{|\mathcal{Z}''_{removed}|} - \frac{\sum_{\mathcal{Z}'} z'}{|\mathcal{Z}'|}\right| \leq \frac{|\mathcal{Z}''_{removed}|}{|\mathcal{Z}|} \end{aligned} \tag{51}$$

Finally, by triangle inequality, we have that

$$\begin{aligned} &\left|\sum_{\mathcal{Z}} \frac{z}{|\mathcal{Z}|} - \sum_{\mathcal{Z}'} \frac{z'}{|\mathcal{Z}'|}\right| \leq \left|\sum_{\mathcal{Z}} \frac{z}{|\mathcal{Z}|} - \sum_{\mathcal{Z}''} \frac{z''}{|\mathcal{Z}''|}\right| + \left|\sum_{\mathcal{Z}''} \frac{z''}{|\mathcal{Z}''|} - \sum_{\mathcal{Z}'} \frac{z'}{|\mathcal{Z}'|}\right| \\ &\leq \frac{|\mathcal{Z}_{added}|}{|\mathcal{Z}|} + \frac{|\mathcal{Z}''_{removed}|}{|\mathcal{Z}|} \leq \frac{|\mathcal{Z}_{added}|}{|\mathcal{Z}|} + \frac{|\mathcal{Z}_{removed}|}{|\mathcal{Z}|} \leq \frac{k}{m} \end{aligned} \tag{52}$$

as desired. A similar argument can be made for the case of $|\mathcal{Z}| < |\mathcal{Z}'|$. □

We now give the classification lemmas:

Lemma C.4. *Given a finite hypothesis class $\mathcal{G} \subseteq \mathcal{X} \to \{0, 1\}$, and two datasets (multisets) $A, B \subset \mathcal{X}$, let:*

$$\begin{aligned} g' &:= \arg\min_{g \in \mathcal{G}} \mathcal{L}(g) \\ \mathcal{L}(g) &:= \frac{1}{|A|}\sum_{a \in A}(1 - g(a)) \; + \; \frac{1}{|B|}\sum_{b \in B} g(b). \end{aligned} \tag{53}$$

Let $n = \min(|A|, |B|)$, and assume that A and B are constructed as follows:

- $A' \sim \mathcal{D}_A^{|A'|}$
- $B' \sim \mathcal{D}_B^{|B'|}$
- *At most a total of m arbitrary (non-i.i.d.) samples are either added to or removed from A' or B', or moved from A' to B' or vice-versa, to create A and B.*

Then, if

$$n \geq 8m \text{ and } n \geq \frac{128}{7} \ln(2|\mathcal{G}|/\delta) \tag{54}$$

then with probability at least $1-\delta$*,*

- *If* $\mathcal{D}_A = \mathcal{D}_B$*, then* $\mathcal{L}(g') > 1/2$
- *Conversely, if* $\mathcal{D}_A$ *and* $\mathcal{D}_B$ *have disjoint support, such that some* $g^* \in \mathcal{G}$ *maps all elements in the support of* $\mathcal{D}_A$ *to 1 and all elements in the support of* $\mathcal{D}_B$ *to 0, then* $\mathcal{L}(g') \leq 1/2$*.*

Proof. Define

$$\mathcal{L}_{clean}(g) := \frac{1}{|A'|} \sum_{a \in A'} (1 - g(a)) + \frac{1}{|B'|} \sum_{b \in B'} g(b). \tag{55}$$

Then, fix any $g \in \mathcal{G}$. From some algebra (see Proposition C.3), it can be shown that

$$\mathcal{L}_{clean}(g) - \frac{2m}{n} \leq \mathcal{L}(g) \leq \mathcal{L}_{clean}(g) + \frac{2m}{n} \tag{56}$$

Now, note that $\mathcal{L}_{clean}$ is the sum of $|A'|$ random variables bounded on $[0, 1/|A'|]$, and $|B'|$ random variables bounded on $[0, 1/|B'|]$, all of which are i.i.d. Then, by Hoeffding's Lemma and Equation 56, with probability $1 - \delta/|\mathcal{G}|$:

$$\begin{aligned} &\mathbb{E}[\mathcal{L}_{clean}(g)] - \sqrt{\left(\frac{1}{|A'|} + \frac{1}{|B'|}\right) \ln(2|\mathcal{G}|/\delta)/2} - \frac{2m}{n} < \mathcal{L}_{clean}(g) - \frac{2m}{n} \leq \mathcal{L}(g) \\ &\leq \mathcal{L}_{clean}(g) + \frac{2m}{n} < \mathbb{E}[\mathcal{L}_{clean}(g)] + \sqrt{\left(\frac{1}{|A'|} + \frac{1}{|B'|}\right) \ln(2|\mathcal{G}|/\delta)/2} + \frac{2m}{n} \end{aligned} \tag{57}$$

Because $|A'|, |B'| \geq n - m$, and applying union bound over all $g \in \mathcal{G}$, we have, with probability $1 - \delta$:

$$\forall g \in \mathcal{G},\ \mathbb{E}[\mathcal{L}_{clean}(g)] - \sqrt{\frac{\ln(2|\mathcal{G}|/\delta)}{n-m}} - \frac{2m}{n} < \mathcal{L}(g) < \mathbb{E}[\mathcal{L}_{clean}(g)] + \sqrt{\frac{\ln(2|\mathcal{G}|/\delta)}{n-m}} + \frac{2m}{n}. \tag{58}$$

Note that:

$$\forall g \in \mathcal{G},\ \mathbb{E}[\mathcal{L}_{clean}(g)] = 1 - \mathbb{E}_{x \in \mathcal{D}_A}[g(x)] + \mathbb{E}_{x \in \mathcal{D}_B}[g(x)]. \tag{59}$$

If $\mathcal{D}_A = \mathcal{D}_B$, then $\forall g \in \mathcal{G},\ \mathbb{E}[\mathcal{L}_{clean}(g)] = 1$, so by Equation 58, we have, with probability $1 - \delta$:

$$\forall g \in \mathcal{G},\ 1 - \sqrt{\frac{\ln(2|\mathcal{G}|/\delta)}{n-m}} - \frac{2m}{n} \leq \mathcal{L}(g), \tag{60}$$

and in particular:

$$1 - \sqrt{\frac{\ln(2|\mathcal{G}|/\delta)}{n-m}} - \frac{2m}{n} < \mathcal{L}(g'). \tag{61}$$

Conversely, if $\mathcal{D}_A$ and $\mathcal{D}_B$ have disjoint support, such that some $g^* \in \mathcal{G}$ maps all elements in the support of $\mathcal{D}_A$ to 1 and all elements in the support of $\mathcal{D}_B$ to 0, then we have:

$$\mathbb{E}[\mathcal{L}_{clean}(g^*)] = 1 - \mathbb{E}_{x \in \mathcal{D}_A}[g^*(x)] + \mathbb{E}_{x \in \mathcal{D}_B}[g^*(x)] = 1 - 1 - 0 = 0. \tag{62}$$

Then, with probability at least $1 - \delta$:

$$\mathcal{L}(g') \leq \mathcal{L}(g^*) < \sqrt{\frac{\ln(2|\mathcal{G}|/\delta)}{n-m}} + \frac{2m}{n}. \tag{63}$$

To complete the proof, we only need to show that

$$\sqrt{\frac{\ln(2|\mathcal{G}|/\delta)}{n-m}} + \frac{2m}{n} \leq \frac{1}{2}. \tag{64}$$

With $m \leq n/8$, this condition becomes:

$$\sqrt{\frac{8\ln(2|\mathcal{G}|/\delta)}{7n}} \leq \frac{1}{4}. \tag{65}$$

or

$$n \geq \frac{128}{7}\ln(2|\mathcal{G}|/\delta) \tag{66}$$

□

Lemma C.5. *Given a finite hypothesis class $\Phi \subset \mathcal{X} \to \mathbb{N}$, and N datasets (multisets) $D_1, D_2, ...D_N \subset \mathcal{X}$, let:*

$$\begin{aligned} \phi' &:= \arg\min_{\phi\in\Phi} \mathcal{L}(\phi) \\ \mathcal{L}(\phi) &:= \sum_{i=1}^{N}\left[\frac{1}{|D_i|}\sum_{x\in D_i}(1 - \mathbb{1}_{(\phi(x)=i)})\right] \end{aligned} \tag{67}$$

Let $n = \min_i(|D_i|)$, and assume that each D_i is constructed as follows:

- $\forall i,\ \ D'_i \sim \mathcal{D}_i^{|D'_i|}$
- *At most a total of m arbitrary (non-i.i.d.) samples are arbitrarily moved between the datasets D'_i, to create the datasets D_i.*

Additionally, assume that $\exists\phi^* \in \Phi : \forall i,\ x \sim \mathcal{D}_i \implies \phi^*(x) = i$. Then, if

$$n \geq \frac{8m}{\epsilon} \text{ and } n \geq \frac{64N\ln(2|\Phi|/\delta)}{7\epsilon^2} \tag{68}$$

then with probability at least $1-\delta$,

$$\forall i \in [N],\ \ \Pr_{x\sim\mathcal{D}_i}(\phi'(x) = i) \geq 1 - \epsilon. \tag{69}$$

Proof. Define

$$\mathcal{L}_{clean}(\phi) := \sum_{i=1}^{N}\left[\frac{1}{|D'_i|}\sum_{x\in D'_i}(1 - \mathbb{1}_{(\phi(x)=i)})\right]. \tag{70}$$

Then, fix any $\phi \in \Phi$. From Proposition C.3 (regarding each transfer of a sample as removing a sample into one multiset D'_i, and inserting a new sample into another) we see that:

$$\mathcal{L}_{clean}(\phi) - \frac{2m}{n} \leq \mathcal{L}(\phi) \leq \mathcal{L}_{clean}(\phi) + \frac{2m}{n} \tag{71}$$

Now, note that $\mathcal{L}_{clean}$ is the sum of $|D'_1|$ random variables bounded on $[0, 1/|D'_1|]$, and $|D'_2|$ random variables bounded on $[0, 1/|D'_2|]$, et cetera, all of which are i.i.d. Then, by Hoeffding's Lemma and Equation 71, with probability $1 - \delta/|\Phi|$:

$$\begin{aligned} &\mathbb{E}[\mathcal{L}_{clean}(\phi)] - \sqrt{\left(\sum_{i\in[N]}\frac{1}{|D'_i|}\right)\ln(2|\Phi|/\delta)/2} - \frac{2m}{n} < \mathcal{L}_{clean}(\phi) - \frac{2m}{n} \leq \mathcal{L}(\phi) \\ &\leq \mathcal{L}_{clean}(\phi) + \frac{2m}{n} < \mathbb{E}[\mathcal{L}_{clean}(\phi)] + \sqrt{\left(\sum_{i\in[N]}\frac{1}{|D'_i|}\right)\ln(2|\Phi|/\delta)/2} + \frac{2m}{n} \end{aligned} \tag{72}$$

Because $\forall i, |D'_i| \geq n - m$, and applying union bound over all $\phi \in \Phi$, we have, with probability $1 - \delta$, $\forall \phi \in \Phi$, :

$$\mathbb{E}[\mathcal{L}_{clean}(\phi)] - \sqrt{\frac{N \ln(2|\Phi|/\delta)}{2(n-m)}} - \frac{2m}{n} < \mathcal{L}(\phi) < \mathbb{E}[\mathcal{L}_{clean}(\phi)] + \sqrt{\frac{N \ln(2|\Phi|/\delta)}{2(n-m)}} + \frac{2m}{n}. \quad (73)$$

Then we have:

$$\begin{aligned} \mathbb{E}[\mathcal{L}_{clean}(\phi')] < \mathcal{L}(\phi') + \sqrt{\frac{N \ln(2|\Phi|/\delta)}{2(n-m)}} + \frac{2m}{n} \leq \\ \mathcal{L}(\phi^*) + \sqrt{\frac{N \ln(2|\Phi|/\delta)}{2(n-m)}} + \frac{2m}{n} < \mathbb{E}[\mathcal{L}_{clean}(\phi^*)] + \sqrt{\frac{2N \ln(2|\Phi|/\delta)}{(n-m)}} + \frac{4m}{n} = \\ \sqrt{\frac{2N \ln(2|\Phi|/\delta)}{(n-m)}} + \frac{4m}{n}. \end{aligned} \quad (74)$$

Where we use the fact that, by the definition of ϕ' as a minimizer, $\mathcal{L}(\phi') \leq \mathcal{L}(\phi^*)$, as well as the fact that, by definition, $\mathcal{L}_{clean}(\phi^*) = 0$.

Also, note that by the definition of $\mathcal{L}_{clean}$, we have that, for any ϕ,

$$\mathbb{E}[\mathcal{L}_{clean}(\phi)] = \sum_{i \in [N]} 1 - \Pr_{x \sim \mathcal{D}_i}(\phi(x) = i) \quad (75)$$

Then, for any particular $i \in [N]$, we have that $1 - \Pr_{x \sim \mathcal{D}_i}(\phi(x) = i) \leq \mathbb{E}[\mathcal{L}_{clean}(\phi)]$. Then, by Equation 74, we have, $\forall i \in [N]$:

$$1 - \Pr_{x \sim \mathcal{D}_i}(\phi'(x) = i) < \sqrt{\frac{2N \ln(2|\Phi|/\delta)}{(n-m)}} + \frac{4m}{n}. \quad (76)$$

By algebra, our desired result (Equation 69) holds as long as:

$$\sqrt{\frac{2N \ln(2|\Phi|/\delta)}{(n-m)}} + \frac{4m}{n} \leq \epsilon \quad (77)$$

Which follows from the given conditions on n. □

C.3 Main Proof of Theorem 3.1

Here, we present the proof of Theorem 3.1. We first split out correctness proof of the main recursive step of the algorithm as a lemma:

Lemma C.6. *In Algorithm 2, suppose that the ground-truth data coverage assumptions given in Equations 4,5, and 6 all hold. Additionally, assume that the relative coverage lower-bound η can be written in the form*

$$\eta = \frac{e^{-n_\Xi \alpha/8}}{1 + e^{-n_\Xi \alpha/8}} \quad (78)$$

for some non-negative integer n_Ξ. Further, assume that for each $s^* \in \mathcal{S}^*_h$, there exists some $s \in \mathcal{S}_h$ that represents approximately the same set of observations. In particular, each index in $[|\tau_A|]$ appears in at most one set $D_{A,h}(s)$ for some s (and likewise for $[|\tau_B|]$ and $D_{B,h}(s)$), and there exists some bijective mapping $\sigma_h : \mathcal{S}_h \to \mathcal{S}^*_h$, such that for most indices j in $[|\tau_A|]$

$$j \in D_{A,h}(\sigma_h^{-1}(\phi^*_h((\tau_A)_h[j]))) \quad (79)$$

and for most indices j in $[|\tau_B|]$

$$j \in D_{B,h}(\sigma_h^{-1}(\phi_h^*((\tau_B)_h[j]))), \tag{80}$$

with at most a combined $\beta(|\tau_A| + |\tau_B|)$ indices in either dataset for which this does not hold.

For any ϵ such that:

$$\epsilon < \frac{\nu}{8} - \beta, \tag{81}$$

and

$$\epsilon + \beta \leq q_{thresh.} < \frac{\nu(1-\epsilon)}{2} - \beta, \tag{82}$$

where $q_{thresh.}$ is the threshold defined on Line 11 of the algorithm, assume that $\forall s_h^*, s_{h+1}^*$, such that s_h^* can transition to s_{h+1}^*,

$$|\mathcal{D}^*(s_h^*, s_{h+1}^*)| \geq \frac{12800(\ln(4|\Phi|^2/\delta) + N_s^2 \ln(n_\Xi + 1)) \ln^2(1/\eta)}{\nu\epsilon^2\eta^2\alpha^4}. \tag{83}$$

Then, with high probability, for each $s^* \in \mathcal{S}_{h+1}^*$, there exists some $s \in \mathcal{S}_{h+1}$ that represents approximately the same set of observations. In particular, each index in $[|\tau_A|]$ appears in at most one set $D_A(s)$ for some s (and likewise for $[|\tau_B|]$ and $D_B(s)$), and there exists some bijective mapping $\sigma_{h+1} : \mathcal{S}_{h+1} \to \mathcal{S}_{h+1}^*$, such that for most indices j in $[|\tau_A|]$

$$j \in D_{A,h+1}(\sigma_{h+1}^{-1}(\phi_{h+1}^*((\tau_A)_{h+1}[j]))) \tag{84}$$

and for most indices j in $[|\tau_A|]$

$$j \in D_{B,h+1}(\sigma_{h+1}^{-1}(\phi_{h+1}^*((\tau_B)_{h+1}[j]))), \tag{85}$$

with at most a combined $(\beta + \epsilon)(|\tau_A| + |\tau_B|)$ indices in either dataset for which this does not hold.

Proof. We first show that the datasets of observation pairs $\mathcal{D}_{new\_pairs}$ defined in Line 21 of the algorithm each correspond uniquely to a pair of ground truth latent states in $\mathcal{S}_h^* \times \mathcal{S}_{h+1}^*$, such that no pair of observations is included in more than one such $\mathcal{D}_{new\_pairs}$ sets, and, with high probability, each pair of observations x, x' is included in the correct $\mathcal{D}_{new\_pairs}$ corresponding to $(\phi_h^*(x), \phi_{h+1}^*(x'))$, with up to at most $(\beta + \epsilon)(|\tau_A| + |\tau_B|)$ exceptions.

For any $s_{pred}^* \in \mathcal{S}_h^*$, consider any two distinct $s^*, s'^* \in \mathcal{S}_{h+1}^*$, such that $|\mathcal{D}^*(s_{pred}^*, s^*)|, |\mathcal{D}^*(s_{pred}^*, s'^*)| > 0$.

Recall the assumption that, without loss of generality,

$$e^\alpha \cdot \frac{\pi_B^{emp.}(s'^*|s_{pred}^*)}{\pi_B^{emp.}(s^*|s_{pred}^*)} \leq \frac{\pi_A^{emp.}(s'^*|s_{pred}^*)}{\pi_A^{emp.}(s^*|s_{pred}^*)}, \tag{86}$$

Multiplying both sides by $\pi_A^{emp.}(s^*|s_{pred}^*)/\pi_B^{emp.}(s'^*|s_{pred}^*)$ yields

$$e^\alpha \cdot \frac{\pi_A^{emp.}(s^*|s_{pred}^*)}{\pi_B^{emp.}(s^*|s_{pred}^*)} \leq \frac{\pi_A^{emp.}(s'^*|s_{pred}^*)}{\pi_B^{emp.}(s'^*|s_{pred}^*)}, \tag{87}$$

From the definition of $\pi^{emp.}$, this is:

$$e^\alpha \cdot \frac{|\mathcal{D}_A^*(s_{pred}^*, s^*)|/|\mathcal{D}_A^*(s_{pred}^*)|}{|\mathcal{D}_B^*(s_{pred}^*, s^*)|/|\mathcal{D}_B^*(s_{pred}^*)|} \leq \frac{|\mathcal{D}_A^*(s_{pred}^*, s'^*)|/|\mathcal{D}_A^*(s_{pred}^*)|}{|\mathcal{D}_B^*(s_{pred}^*, s'^*)|/|\mathcal{D}_B^*(s_{pred}^*)|}. \tag{88}$$

Multiplying both sides by $|\mathcal{D}_A^*(s_{pred}^*)|/|\mathcal{D}_B^*(s_{pred}^*)|$ and taking the logarithms yields:

$$\alpha + \ln\left(\frac{|\mathcal{D}_A^*(s_{pred}^*, s^*)|}{|\mathcal{D}_B^*(s_{pred}^*, s^*)|}\right) \leq \ln\left(\frac{|\mathcal{D}_A^*(s_{pred}^*, s'^*)|}{|\mathcal{D}_B^*(s_{pred}^*, s'^*)|}\right). \tag{89}$$

By the definitions of $\underline{c}_{s_h^*,s_{h+1}^*}$ and $\bar{c}_{s_h^*,s_{h+1}^*}$ in Corollary C.2, and the fact that $\xi = \alpha/4$, we see that there must be at least two values in Ξ between $\bar{c}_{s^*_{pred},s^*}$ and $\underline{c}_{s^*_{pred},s'^*}$; that is to say:

$$\underline{c}_{s^*_{pred},s^*} \leq \bar{c}_{s^*_{pred},s^*} < \bar{c}_{s^*_{pred},s^*} + \xi < \underline{c}_{s^*_{pred},s'^*} - \xi < \underline{c}_{s^*_{pred},s'^*} \leq \bar{c}_{s^*_{pred},s'^*} \tag{90}$$

Therefore, by Corollary C.2 we have, with probability at least $1 - \delta/2$, for any s^* such that $|\mathcal{D}^*(s^*_{pred}, s^*)| > 0$:

- At least $(1-\epsilon)|\mathcal{D}^*(s^*_{pred}, s^*)|$ of the samples in $\mathcal{D}^*(s^*_{pred}, s^*)$ will be mapped by f to $\underline{c}_{s^*_{pred},s^*}$ or $\bar{c}_{s^*_{pred},s^*}$
- for some choice of $\hat{c}_{s^*_{pred},s^*} \in \{\underline{c}_{s^*_{pred},s^*}, \bar{c}_{s^*_{pred},s^*}\}$, at least $(1-\epsilon)/2 \cdot |\mathcal{D}^*(s^*_{pred}, s^*)|$ of the samples in $\mathcal{D}^*(s^*_{pred}, s^*)$ will be mapped to $\hat{c}_{s^*_{pred},s^*}$.
- By definition, $\{\underline{c}_{s^*_{pred},s^*}, \bar{c}_{s^*_{pred},s^*}\} \subset \{\hat{c}_{s^*_{pred},s^*} - \xi, \hat{c}_{s^*_{pred},s^*}, \hat{c}_{s^*_{pred},s^*} + \xi\}$.
- Furthermore, for no two states s^*, s'^*, with $\hat{c}_{s^*_{pred},s^*} \in \{\underline{c}_{s^*_{pred},s^*}, \bar{c}_{s^*_{pred},s^*}\}$ and $\hat{c}_{s^*_{pred},s'^*} \in \{\underline{c}_{s^*_{pred},s'^*}, \bar{c}_{s^*_{pred},s'^*}\}$ chosen arbitrarily, will the sets $\{\hat{c}_{s^*_{pred},s^*} - \xi, \hat{c}_{s^*_{pred},s^*}, \hat{c}_{s^*_{pred},s^*} + \xi\}$ and $\{\hat{c}_{s^*_{pred},s'^*} - \xi, \hat{c}_{s^*_{pred},s'^*}, \hat{c}_{s^*_{pred},s'^*} + \xi\}$ overlap (By Equation 90).
- Recall that by assumption, $|\mathcal{D}^*(s^*_{pred}, s^*)| \geq \nu(|\tau_A| + |\tau_B|)$. Therefore, at least $(\nu(1-\epsilon)/2)(|\tau_A| + |\tau_B|)$ of the samples in $\mathcal{D}^*(s^*_{pred}, s^*)$ will be mapped to $\hat{c}_{s^*_{pred},s^*}$.
- The total number of samples in $\mathcal{D}^*(s^*_{pred}, s'^*)$, over *all* choices of s'^*, which are *not* mapped by f to a value in the respective set $\{\hat{c}_{s^*_{pred},s'^*} - \xi, \hat{c}_{s^*_{pred},s'^*}, \hat{c}_{s^*_{pred},s'^*} + \xi\}$, is at most $\epsilon|\mathcal{D}^*(s^*_{pred})|$.
- $\epsilon|\mathcal{D}^*(s^*_{pred})| \leq \epsilon(|\tau_A| + |\tau_B|)$.

Therefore, as long as $(\nu(1-\epsilon)/2) > \epsilon$, then among the pairs in $\mathcal{D}^*(s^*_{pred}, \cdot) := \uplus_{s'^*} \mathcal{D}^*(s^*_{pred}, s'^*)$, if there is any $z \in \Xi$ such that $> \epsilon(|\tau_A| + |\tau_B|)$ of the pairs are mapped by f to z, then we know that the set of elements in $\mathcal{D}^*(s^*_{pred}, \cdot)$ which are mapped to $\{z-1, z, z+1\}$ contains at least $(1-\epsilon)$ of the elements of the set $\mathcal{D}^*(s^*_{pred}, s^*)$ for some s^*; furthermore, such a z exists for each possible value of s^* where $|\mathcal{D}^*(s^*_{pred}, s^*)| > 0$, and, for distinct s^* and s'^*, these values ($\{z-\xi, z, z+\xi\}$ and $\{z'-\xi, z', z'+\xi\}$) are non-overlapping. Consequently, by identifying subsets of $\mathcal{D}^*(s^*_{pred}, \cdot)$ of size greater than $\epsilon(|\tau_A| + |\tau_B|)$ that f maps to the same value, and expanding these subsets to the elements in $\mathcal{D}^*(s^*_{pred}, \cdot)$ mapped to adjacent values in Ξ, we can partition $\mathcal{D}^*(s^*_{pred}, \cdot)$ into subsets corresponding to each $\mathcal{D}^*(s^*_{pred}, s^*)$, with at most $\epsilon|\mathcal{D}^*(s^*_{pred}, \cdot)|$ errors.

Note however that we do not have access to $\mathcal{D}^*(s^*_{pred}, \cdot)$, only to pairs$(s_{pred})$ (where $\sigma_h(s_{pred}) = s^*_{pred}$). However, by assumption, $\mathcal{D}^*(s^*_{pred}, \cdot)$ and pairs(s_{pred}) differ (in terms of symmetric difference) by at most $\beta(|\tau_A| + |\tau_B|)$. Therefore, we claim that, if

$$\epsilon + \beta \leq q_{thresh.} < \frac{\nu(1-\epsilon)}{2} - \beta \tag{91}$$

then, we can identify values of $\hat{c}_{s^*_{pred},s'^*}$ (for some s'^*) as those values $j\xi - \frac{n_\Xi \xi}{2}$ for which (as shown in Line 20 of Algorithm 2):

$$\text{pred\_succ}[j] > q_{thresh.}(|\tau_A| + |\tau_B|), \tag{92}$$

and, conversely, if

$$\text{pred\_succ}[j] \leq q_{thresh.}(|\tau_A| + |\tau_B|), \tag{93}$$

then $j\xi - \frac{n_\Xi \xi}{2}$ does not correspond to some $\bar{c}_{s^*_{pred},s'^*}$ or $\underline{c}_{s^*_{pred},s'^*}$.

To validate this claim, note that if Equation 91 holds, then:

$$\begin{aligned}
&\text{\# of samples } (x,x') \text{ in pairs}(s_{pred}) \text{ such that } f(x,x')=z, \text{ if } \nexists s^*: z \in \{\bar{c}_{s^*_{pred},s^*}, \underline{c}_{s^*_{pred},s^*}\} \leq \\
&\quad \text{\# of samples } (x,x') \text{ in } \mathcal{D}^*(s^*_{pred}) \text{ such that } f(x,x')=z, \text{ if } \nexists s^*: z \in \{\bar{c}_{s^*_{pred},s^*}, \underline{c}_{s^*_{pred},s^*}\} \\
&\qquad + |\text{pairs}(s_{pred}) \setminus \mathcal{D}^*(s^*_{pred})| \leq \\
&\qquad \epsilon(|\tau_A|+|\tau_B|) + |\text{pairs}(s_{pred}) \setminus \mathcal{D}^*(s^*_{pred})| \leq \\
&\qquad \textbf{(Note this line:)} \quad \epsilon(|\tau_A|+|\tau_B|) + \beta(|\tau_A|+|\tau_B|) \leq \\
&\qquad q_{thresh.}(|\tau_A|+|\tau_B|) < \\
&\qquad (\nu(1-\epsilon)/2)(|\tau_A|+|\tau_B|) - \beta(|\tau_A|+|\tau_B|) \leq \\
&\qquad (\nu(1-\epsilon)/2)(|\tau_A|+|\tau_B|) - |\mathcal{D}^*(s^*_{pred}) \setminus \text{pairs}(s_{pred})| \leq \\
&\quad \text{\# of samples } (x,x') \text{ in } \mathcal{D}^*(s^*_{pred}) \text{ such that } f(x,x')=z, \text{ if } \exists s^*: z \in \{\bar{c}_{s^*_{pred},s^*}, \underline{c}_{s^*_{pred},s^*}\} \\
&\qquad - |\mathcal{D}^*(s^*_{pred}) \setminus \text{pairs}(s_{pred})| \leq \\
&\text{\# of samples } (x,x') \text{ in pairs}(s_{pred}) \text{ such that } f(x,x')=z, \text{ if } \exists s^*: z \in \{\bar{c}_{s^*_{pred},s^*}, \underline{c}_{s^*_{pred},s^*}\}
\end{aligned}$$

Therefore, for any s^*_{pred}, we can define:

$$\mathcal{D}^*_{new\_pairs}(s^*_{pred}, j, j') := \{(x,x') | (x,x') \in \biguplus_{k=j-1}^{j'} \text{pred\_succ}^*[k]\} \tag{94}$$

where

$$\text{pred\_succ}^*[k] := \{(x_h, x_{h+1}) \in \mathcal{D}^*(s^*_{pred}, \cdot) | f(x_h, x_{h+1}) = k\xi - \frac{n_\Xi \xi}{2}\}. \tag{95}$$

If j and j' are chosen as in Line 19 and 20 of Algorithm 2, then for any pair (s^*_{pred}, s^*) there is a unique set $\mathcal{D}^*_{new\_pairs}(s^*_{pred}, j, j')$ containing at least a $(1-\epsilon)$ fraction of the samples in $\mathcal{D}(s^*_{pred}, s^*)$. Furthermore, note that by assumption, summing over *all pairs* (s^*_{pred}, s^*), the sets $\mathcal{D}^*_{new\_pairs}(s^*_{pred}, j, j')$ and $\mathcal{D}_{new\_pairs}$ can differ by at most $\beta(|\tau_A|+|\tau_B|)$ members in total (because all datasets $\mathcal{D}^*(s^*_{pred}, \cdot)$ and pairs(s_{pred}) differ by at most this many members in total). Therefore, we have shown that, with high probability, each pair of observations x, x' is included in the correct $\mathcal{D}_{new\_pairs}$ corresponding to $(\phi^*_h(x), \phi^*_{h+1}(x'))$, with up to at most $(\beta+\epsilon)(|\tau_A|+|\tau_B|)$ exceptions. (Furthermore, different sets $\mathcal{D}_{new\_pairs}$ are non-overlapping by construction.) Then we only need to show that, with high probability, Line 26 of Algorithm 2 will only merge two sets $\mathcal{D}_{new}$ if these sets correspond to the same latent state. By Lemma C.4, taking a union bound over all pairs of latent-state sequential latent-state pairs, we have, with probability at least $1-\delta/2$ that, if

$$\nu \geq 8(\beta+\epsilon) \text{ and } |\mathcal{D}^*(s^*_h, s^*_{h+1})| \geq \frac{128}{7} \ln(4 \cdot (\text{\# of latent state pairs}) \cdot |\mathcal{G}_{h+1}|/\delta) \tag{96}$$

then the classifiers trained in Line 25 will have loss greater than 1/2 if and only if the two datasets being compared correspond to the same latent state. Also note that (# of latent state pairs) $\leq 1/\nu$; then, under the assumption that $|\mathcal{G}_{h+1}| \leq |\Phi| (\leq |\Phi|^2)$, we have

$$\frac{128}{7} \ln(4 \cdot (\text{\# of latent state pairs}) \cdot |\mathcal{G}_{h+1}|/\delta) \leq \frac{128}{7} (\ln(4|\Phi^2|/\delta) + \ln(1/\nu)). \tag{97}$$

Note that by Equation 40 (and $\xi = \alpha/4$), we have

$$\frac{16 \ln(1/\eta)}{\eta \alpha^2} \geq 1, \tag{98}$$

so we can write:

$$\frac{128}{7} \ln(4 \cdot (\text{\# of latent state pairs}) \cdot |\mathcal{G}_{h+1}|/\delta) \leq \frac{128 \cdot 256}{7} \frac{(\ln(4|\Phi^2|/\delta) + \ln(1/\nu)) \ln^2(1/\eta)}{\eta^2 \alpha^4}. \tag{99}$$

Also noting that $\epsilon \leq 1$, and $\ln(4|\Phi|^2/\delta) \geq \ln(4) \geq 1$, and $1/\nu \geq \{\ln(1/\nu), 1\}$, we have:

$$\frac{128}{7}\ln(4 \cdot (\#\text{ of latent state pairs}) \cdot |\mathcal{G}_{h+1}|/\delta) \leq \frac{128 \cdot 256}{7}\frac{(\ln(4|\Phi^2|/\delta) + \ln(4|\Phi^2|/\delta))\ln^2(1/\eta)}{\nu\epsilon^2\eta^2\alpha^4}. \tag{100}$$

Then, because $N_s^2 \ln(n_\Xi + 1) > 0$, and $128 \cdot 256 \cdot 2/7 \leq 12800$, we have

$$\frac{128}{7}\ln(4 \cdot (\#\text{ of latent state pairs}) \cdot |\mathcal{G}_{h+1}|/\delta) \leq \frac{12800(\ln(4|\Phi^2|/\delta) + N_s^2\ln(n_\Xi + 1))\ln^2(1/\eta)}{\nu\epsilon^2\eta^2\alpha^4}. \tag{101}$$

Therefore, the number of observations of each latent state pair $|\mathcal{D}^*(s_h^*, s_{h+1}^*)|$ assumed in Equation 83 is sufficient to ensure that all datasets $\mathcal{D}_{new}$ will be merged correctly. Then, by union bound, with probability at least $1-\delta$, the samples corresponding to the indices in $D_{A,h+1}(s)$ and $D_{B,h+1}(s)$ will correspond to the observations of a unique latent state s^*, with up to at most $(\beta+\epsilon)(|\tau_A|+|\tau_B|)$ exceptions. □

The following lemma is essentially Theorem 3.1, with a minor additional assumption:

Lemma C.7. *Assume that Algorithm 2 is given datasets τ_A and τ_B such that the assumptions given in Equations 1, 4,5, and 6 all hold. Additionally, assume that the relative coverage lower-bound η can be written in the form*

$$\eta = \frac{e^{-n_\Xi\alpha/8}}{1 + e^{-n_\Xi\alpha/8}} \tag{102}$$

for some non-negative integer n_Ξ. Then, for any given $\delta, \epsilon_0 \geq 0$, if $\forall s_h^*, s_{h+1}^*$, such that s_h^* can transition to s_{h+1}^*,

$$|\mathcal{D}^*(s_h^*, s_{h+1}^*)| \geq \frac{819200H^2(\ln(8H|\Phi|^2/\delta) + N_s^2\ln(n_\Xi+1))\ln^2(1/\eta)}{\nu\eta^2\alpha^4} \cdot \max\left(\frac{1}{\nu^2}, \frac{1}{\epsilon_0^2\nu'^2}\right), \tag{103}$$

then, with probability at least $1-\delta$, the encoders ϕ_h' returned by the algorithm will each have accuracy on at least $1-\epsilon_0$, in the sense that, under some bijective mapping $\sigma_h : \mathcal{S}_h \to \mathcal{S}_h^*$,

$$\forall s^* \in \mathcal{S}_h^*, \quad \Pr_{x \sim \mathcal{Q}(s^*, P_h^e)}(\phi_h'(x) = \sigma_h^{-1}(\phi_h^*(x))) \geq 1 - \epsilon_0. \tag{104}$$

Proof. Note that the conclusion applies at timestep $h = 1$ vacuously: there is only one latent state, and ϕ_1' returns a constant value. Further, $D_{A,1}(s_1)$ and $D_{B,1}(s_1)$ contain exactly the sets of trajectories in τ_A and τ_B which visit s_1^* at step 1.

For subsequent steps, we apply Lemma C.6 recursively, with:

- $\epsilon = \min(\frac{\nu}{8H}, \frac{\nu'\epsilon_0}{8H})$
- $\beta = (h-1)\epsilon$
- $\delta_{\text{Lemma C.6}} := \delta/(2H)$.

Note that the assumptions in Equations 81 and 82 are met, because:

$$\epsilon = \epsilon + \beta - \beta = h\epsilon - \beta < H\epsilon - \beta = H\min(\frac{\nu}{8H}, \frac{\nu'\epsilon_0}{8H}) - \beta < \frac{H\nu}{8H} - \beta < \frac{\nu}{8} - \beta \tag{105}$$

and

$$\epsilon + \beta = h\min\left(\frac{\nu}{8H}, \frac{\nu'\epsilon_0}{8H}\right) \leq \frac{h\nu}{8H} = q_{thresh.} \tag{106}$$

and

$$\begin{aligned} q_{thresh.} = \\ q_{thresh.} + \beta + \frac{\epsilon\nu}{2} - \beta - \frac{\epsilon\nu}{2} = \\ \frac{h\nu}{8H} + (h - 1 + \nu/2)\min\left(\frac{\nu}{8H}, \frac{\nu'\epsilon_0}{8H}\right) - \beta - \frac{\epsilon\nu}{2} \leq \\ \frac{h\nu}{8H} + \frac{(h - 1 + \nu/2)\nu}{8H} - \beta - \frac{\epsilon\nu}{2} < \\ \frac{2H\nu}{8H} - \frac{\epsilon\nu}{2} - \beta < \\ \frac{\nu(1-\epsilon)}{2} - \beta. \end{aligned} \tag{107}$$

Also, note that the assumption of Equation 83 is met (by comparison to Equation 103, with $\epsilon = \min(\frac{\nu}{8H}, \frac{\nu'\epsilon_0}{8H})$ and $\delta_{\text{Lemma C.6}} := \delta/(2H)$.) Finally, the inductive hypothesis, that $D_{A,h}(s)$ and $D_{B,h}(s)$ correspond to observations of some state s^*, with at most $\beta(|\tau_A| + |\tau_B|)$ exceptions, can be shown to hold. In particular, at iteration $h \geq 2$, we have that the input dataset has at most $\beta_h(|\tau_A|+|\tau_B|) = (\beta_{h-1}+\epsilon)(|\tau_A|+|\tau_B|)$ errors: we can confirm that $\beta_h = (h-1)\epsilon = (h-2)\epsilon+\epsilon = \beta_{h-1} + \epsilon$ for all $h \geq 2$, with $\beta_1 = 0$ (because there are no errors in $D_{A,1}(s_1)$ and $D_{B,1}(s_1)$).

Therefore, by induction and union bound, we can conclude that, with probability at least $1-\delta/2$, for each $h \in [H]$ and each $s^* \in \mathcal{S}^*_h$, there exists some $s \in \mathcal{S}_h$ that represents approximately the same set of observations. In particular, each index in $[|\tau_A|]$ appears in at most one set $D_A(s)$ for some s (and likewise for $[|\tau_B|]$ and $D_B(s)$), and there exists some bijective mapping $\sigma_h : \mathcal{S}_h \to \mathcal{S}^*_h$, such that for most indices j in $[|\tau_A|]$

$$j \in D_{A,h}(\sigma_h^{-1}((\phi^*(\tau_A)_h[j]))) \tag{108}$$

and for most indices j in $[|\tau_A|]$

$$j \in D_{B,h}(\sigma_h^{-1}(\phi^*((\tau_B)_h[j]))), \tag{109}$$

with at most a combined $(H-1)\min(\frac{\nu}{8H}, \frac{\nu'\epsilon_0}{8H})(|\tau_A|+|\tau_B|)$ indices in either dataset for which this does not hold. Note in particular that fewer than $\frac{\nu'\epsilon_0}{8}(|\tau_A|+|\tau_B|)$ indices will be mis-categorized at any timestep. Then, by application of Lemma C.5 with $n \geq \nu'(|\tau_A|+|\tau_B|)$, $m = \frac{\nu'\epsilon_0}{8}(|\tau_A|+|\tau_B|)$, $\delta_{\text{Lemma C.5}} = \delta/(2H)$ and $\epsilon = \epsilon_0$, we have that as long as:

$$\nu'(|\tau_A| + |\tau_B|) \geq \frac{64 \cdot \max_i |\mathcal{S}_i| \cdot \ln(4H|\Phi|/\delta)}{7\epsilon_0^2}, \tag{110}$$

then, by union bound, with probability at least $1-\delta$, the encoders learned on line 46 of Algorithm 2 will each have accuracy at least $1-\epsilon_0$ as in Equation 104, as desired. All that remains to be shown is that Equation 110 holds. Note that this equation can be re-written as:

$$|\tau_A| + |\tau_B| \geq \frac{64 \cdot \max_h |\mathcal{S}_h| \cdot \ln(4H|\Phi|/\delta)}{7\nu'\epsilon_0^2}. \tag{111}$$

Then, we have:

$$\begin{aligned}
&|\tau_A| + |\tau_B| \geq \\
&|\mathcal{D}^*(s_h^*, s_{h+1}^*)| \geq \\
&\frac{819200H^2(\ln(8H|\Phi|^2/\delta) + N_s^2\ln(n_\Xi+1))\ln^2(1/\eta)}{\nu\eta^2\alpha^4} \cdot \max\left(\frac{1}{\nu^2}, \frac{1}{\epsilon_0^2\nu'^2}\right) \geq \\
&\frac{819200H^2(\ln(8H|\Phi|^2/\delta) + N_s^2\ln(n_\Xi+1))\ln^2(1/\eta)}{\epsilon_0^2\nu'^2\nu\eta^2\alpha^4} \geq \\
\text{(by Equation 98)}\quad &\frac{3200H^2(\ln(8H|\Phi|^2/\delta) + N_s^2\ln(n_\Xi+1))}{\epsilon_0^2\nu'^2\nu} \geq \\
\text{(log. of integer} \geq 0)\quad &\frac{3200H^2\ln(8H|\Phi|^2/\delta)}{\epsilon_0^2\nu'^2\nu} \geq \\
\text{(By definition, } (1/\nu' \geq \max_h |\mathcal{S}_h|)\quad &\frac{3200H^2\max_h|\mathcal{S}_h|\ln(8H|\Phi|^2/\delta)}{\epsilon_0^2\nu'\nu} \geq \\
&\frac{64\cdot\max_h|\mathcal{S}_h|\cdot\ln(4H|\Phi|/\delta)}{7\nu'\epsilon_0^2}
\end{aligned} \tag{112}$$

completing the proof. □

Finally, we prove Theorem 3.1:

Theorem 3.1. *Assume that CRAFT (Algorithm 2 in the Appendix) is given datasets τ_A and τ_B such that the assumptions given in Equations 1, 4,5, and 6 all hold. Then there exists an*

$$f\left(H, |\Phi|, N_s, \frac{1}{\delta}, \frac{1}{\epsilon_0}, \frac{1}{\nu}, \frac{1}{\nu'}, \frac{1}{\eta}, \frac{1}{\alpha}\right) \in \mathcal{O}^*\left(\frac{H^2(\ln(|\Phi|/\delta) + N_s^2)}{\nu\eta^2\alpha^4} \cdot \max\left(\frac{1}{\nu^2}, \frac{1}{\epsilon_0^2\nu'^2}\right)\right), \tag{16}$$

where $\mathcal{O}^*(f(x)) := \mathcal{O}(f(x)\log^k(f(x)))$, such that for any given $\delta, \epsilon_0 \geq 0$, if $\forall s_h^*, s_{h+1}^*$ such that s_h^* can transition to s_{h+1}^*, $|\mathcal{D}^*(s_h^*, s_{h+1}^*)| \geq f\left(H, |\Phi|, N_s, \frac{1}{\delta}, \frac{1}{\epsilon_0}, \frac{1}{\nu}, \frac{1}{\nu'}, \frac{1}{\eta}, \frac{1}{\alpha}\right)$, then, with probability at least $1-\delta$, the encoders ϕ_h' returned by the algorithm will each have accuracy on at least $1-\epsilon_0$, in the sense that, under some bijective mapping $\sigma_h : \mathcal{S}_h \to \mathcal{S}_h^*$,

$$\forall s^* \in \mathcal{S}_h^*, \quad \Pr_{x\sim\mathcal{Q}(s^*, P_h^e)}(\phi_h'(x) = \sigma_h^{-1}(\phi_h^*(x))) \geq 1-\epsilon_0. \tag{17}$$

Proof. This final theorem follows close-to-directly from Lemma C.7, with the caveat that we no longer assume that $\eta = (e^{-n_\Xi\alpha/8})/(1+e^{-n_\Xi\alpha/8})$ for some non-negative integer n_Ξ. To do this, it is important to note that the provided η is a *lower bound*: if we replace η in the algorithm with any arbitrary $\eta' \leq \eta$, then Lemma C.7 will still apply, with a sample-complexity in terms of η' rather than η. Similarly, α is a lower-bound, and so Lemma C.7 will apply for any smaller α'. Our task is then to replace η and α with some η' and α', such that the *asymptotic* sample complexity as given by Equation 16 still applies. For simplicity, we can write the sample-complexity given in Equation 103 as:

$$|\mathcal{D}^*(s_h^*, s_{h+1}^*)| \geq \frac{(C_1 + C_2\ln(n_\Xi+1))\ln^2(1/\eta')}{\eta'^2\alpha'^4}, \tag{113}$$

where C_1 and C_2 are independent of α, η, and n_Ξ. Now, we must choose η' such that:

$$\eta \geq \eta' = \frac{e^{-n_\Xi\alpha/8}}{1+e^{-n_\Xi\alpha/8}}\left(= \frac{1}{1+e^{n_\Xi\alpha/8}}\right). \tag{114}$$

An obvious choice is to take (recalling that by definition, $\eta \leq 1/2$, so $\ln(\eta^{-1}-1) > 0$):

$$n_\Xi := \lceil 8\ln(\eta^{-1}-1)/\alpha \rceil \tag{115}$$

so that:

$$\eta' = \frac{1}{1 + e^{\lceil 8 \ln(\eta^{-1}-1)/\alpha \rceil \alpha/8}} \tag{116}$$

Then we have:

$$\eta \geq \eta' \geq \frac{1}{1 + e^{(8 \ln(\eta^{-1}-1)/\alpha + 1)\alpha/8}} \geq \eta \cdot e^{-\alpha/8} \tag{117}$$

so that we have sufficient samples for Lemma C.7 if:

$$|\mathcal{D}^*(s_h^*, s_{h+1}^*)| \geq \frac{(C_1 + C_2 \ln(8 \ln(\eta^{-1} - 1)/\alpha' + 2))(\ln(1/\eta) + \alpha'/8)^2 e^{\alpha'/4}}{\eta^2 \alpha'^4}, \tag{118}$$

Strictly speaking, Equation 118 with $\alpha' = \alpha$ satisfies the "big-O" asymptotic complexity given in Equation 16 in terms of $1/\alpha$ and $1/\eta$ as these quantities approach infinity. However, if we just take $\alpha' = \alpha$, notice that Equation 118 seems to require an exponentially large number of samples for large α. Recall though that α is a *lower bound*, so we can simply select an arbitrarily lower α' in the case of large α. In particular, if we take $\alpha' = \min(1, \alpha)$, then $\alpha' \leq \alpha$ as needed, and (ignoring lower-order polynomial terms and all logarithmic factors), the dependence of our sample complexity on α becomes:

$$\min(e^{\alpha/4}/\alpha^4, e^{1/4}/1^4) = \min(e^{\alpha/4}/\alpha^4, e^{1/4}) \leq C \cdot 1/\alpha^4 \tag{119}$$

so that the sample complexity is bounded even for large α.

These modifications to η and α are performed on Lines 1-3 of Algorithm 2, so the overall asymptotic sample complexity given in Equation 16 holds for the algorithm overall, with the input α and η. □

D Experiment Details

For the hyperparameters η, ν and α of CRAFT, we use the "population" values based on the ground-truth dynamics and policy. In other words, we set:

$$\begin{aligned} e^{\alpha} = \min_{s_h^*, s_{h+1}^*, s_{h+1}'^*} \max \Bigg[& \left(\frac{\Pr_{\pi_A}(s_{h+1}'^* | s_h^*) / \Pr_{\pi_A}(s_{h+1}^* | s_h^*)}{\Pr_{\pi_B}(s_{h+1}'^* | s_h^*) / \Pr_{\pi_B}(s_{h+1}^* | s_h^*)} \right), \\ & \left(\frac{\Pr_{\pi_B}(s_{h+1}'^* | s_h^*) / \Pr_{\pi_B}(s_{h+1}^* | s_h^*)}{\Pr_{\pi_A}(s_{h+1}'^* | s_h^*) / \Pr_{\pi_A}(s_{h+1}^* | s_h^*)} \right) \Bigg] = \frac{0.75/0.25}{0.5/0.5} = 3 \end{aligned} \tag{120}$$

so $\alpha = \ln(3)$, and

$$\nu = \min_{s_h^*, s_{h+1}^*} \frac{\Pr_{\pi_A}(s_h^*, s_{h+1}^*) + \Pr_{\pi_B}(s_h^*, s_{h+1}^*)}{2} = \frac{1/4 + 1/16}{2} = \frac{5}{32}, \tag{121}$$

and

$$\eta = \min_{s_h^*, s_{h+1}^*} \frac{\Pr_{\pi_B}(s_h^*, s_{h+1}^*)}{\Pr_{\pi_A}(s_h^*, s_{h+1}^*) + \Pr_{\pi_B}(s_h^*, s_{h+1}^*)} = \frac{1/16}{1/4 + 1/16} = \frac{1}{5}. \tag{122}$$

For the "Single observation classification" and "Paired observation classification" baselines, we select the feature ϕ_h (or feature-pair ϕ_h, ϕ_{h+1}) such that the mutual information between $\phi_h(x_h)$ (respectively, $(\phi_h(x_h), \phi_{h+1}(x_{h+1}))$) and the agent's identity is maximized on the collected trajectories.

The "Average Encoder Accuracy" was computed based on the "population" behavior of the environment: that is, the accuracy of encoder ϕ_h' which extracts the feature $(s_h^* \text{ XOR } e_h^i)$ from x_h is computed as $\max(\Pr(e_h^i = 1), \Pr(e_h^i = 0))$, which can be determined analytically from the parameters of Markov chain e^i. For a given algorithm, this quantity was then averaged over timesteps for the returned encoder.

For the "Paired observation classification" baseline, note that for timesteps $h = 2$ through $h = H-1$, the suggested baseline could refer two distinct encoders: the encoder ϕ_h such that $\phi_h(x_h)$ and some

$\phi_{h+1}(x_{h+1})$ are together most informative at predicting the agent observing (x_h, x_{h+1}); *or* the encoder ϕ'_h such that $\phi'_h(x_h)$ and some $\phi_{h-1}(x_{h-1})$ are together most informative at predicting the agent observing (x_{h-1}, x_h). In reporting the final encoder accuracies, took the average accuracy of these two encoders.