

# Selecting Belief-State Approximations in Simulators with Latent States

Nan Jiang

University of Illinois Urbana-Champaign

NANJIANG@ILLINOIS.EDU

## Abstract

State resetting is a fundamental but often overlooked capability of simulators. It supports sample-based planning by allowing resets to previously encountered simulation states, and enables calibration of simulators using real data by resetting to states observed in real-system traces. While often taken for granted, state resetting in complex simulators can be nontrivial: when the simulator comes with latent variables (states), state resetting requires sampling from the posterior over the latent state given the observable history, a.k.a. the belief state (Silver and Veness, 2010). While exact sampling is often infeasible, many approximate belief-state samplers can be constructed, raising the question of how to select among them using only sampling access to the simulator.

In this paper, we show that this problem reduces to a general conditional distribution-selection task and develop a new algorithm and analysis under sampling-only access. Building on this reduction, the belief-state selection problem admits two different formulations: LATENT STATE-BASED SELECTION, which directly targets the conditional distribution of the latent state, and OBSERVATION-BASED SELECTION, which targets the induced distribution over the observation. Interestingly, these formulations differ in how their guarantees interact with the downstream roll-out methods: perhaps surprisingly, OBSERVATION-BASED SELECTION may fail under the most natural roll-out method (which we call SINGLE-RESET) but enjoys guarantees under the less conventional alternative (which we call REPEATED-RESET). Together with discussion on issues such as distribution shift and the choice of sampling policies, our paper reveals a rich landscape of algorithmic choices, theoretical nuances, and open questions, in this seemingly simple problem.

## 1. Introduction

Applying reinforcement learning (RL) to real-world domains often relies on training and evaluating policies in simulation. A basic functionality of simulation is *state resetting/loading*, i.e., setting the simulator into a state that is either previously encountered in simulation or observed in the real system. The former enables sample-based planning—for example, MCTS methods roll-out multiple trajectories from the same state (Kocsis and Szepesvári, 2006; Browne et al., 2012)—while the latter allows one to calibrate the simulator by comparing its predicted next-state to what occurs in reality (Liu et al., 2025).

Despite often taken for granted in research papers, state resetting can be highly nontrivial in complex simulators, especially when they come with *latent variables* that are introduced to model the generative processes of the observables but cannot be measured in the real systems. Naïve approaches, such as loading the saved latent states (e.g., loading previously dumped RAM state (Ecoffet et al., 2019)), is not only infeasible in real systems—since the values of the latent variables are nowhere to be found—but also problematic for resetting to a

previous simulation state; for example, policies trained with such naïve resetting may depend their actions on privileged information in the latent states, and thus may face performance degradation when distilled to a policy that operators only on observable information (Jiang, 2019; Weihs et al., 2021). The correct formulation is to view the simulator as a POMDP, which induces an MDP where the observable *history* (i.e., the sequence of observations and actions) is treated as the state. State resetting amounts to using the observable history to set the values of the latent variables. Mathematically, we should **sample from the posterior distribution of latent variables conditioned on the observable history**, a.k.a. the *belief state* of the POMDP (Silver and Veness, 2010).

While the belief state is conceptually well-defined for any POMDP, exact sampling from belief states can be computationally demanding, especially when the observation space and the latent state space are high-dimensional and the latent dynamics and the emission process are complex black-boxes. To address this challenge, algorithms for approximately sampling from such a distribution have been proposed: for example, the problem can be viewed as an instance of approximate Bayesian computation (ABC), and Silver and Veness (2010) apply rejection sampling to approximate the belief state. However, rejection sampling, when implemented exactly, incurs exponential-in-horizon sample complexity even when the observation space is finite and small, and requires problem-specific heuristics to trade off accuracy for efficiency. Likewise, techniques from related areas such as Simulation-based Inference (SBI) also come with design choices that need to be carefully tuned. This naturally gives rise to the following question:

*Given multiple approximations to the belief state, how can we select from them, and what theoretical guarantees can be obtained?*

In this paper, we explore the multi-faceted nature of this problem. We first show that finding a good belief-state approximation can be reduced to a general conditional distribution-selection problem, and provide a new algorithm and an analysis for the latter under only sampling access to the candidate conditionals (Section 3). Building on this reduction, we then show that belief-state selection itself admits two distinct formulations: **latent state-based selection**, which directly targets the posterior of latent state given history, and **observation-based selection**, which targets the induced observable transition model (Section 4). These two formulations behave differently in the presence of redundant latent variables and, perhaps more importantly, interact in subtle but consequential ways with how we use the selected belief state in downstream tasks. Perhaps surprisingly, we show that, when the selected belief-state approximation is used to estimate Q-values via Monte-Carlo roll-outs, OBSERVATION-BASED SELECTION can have degenerate behavior under the most natural roll-out procedure which we call **Single-Reset**, but enjoy guarantees under the counter-intuitive **Repeated-Reset** roll-out (Section 5; see also Table 1). We conclude the paper with further discussions on the issues related to distribution shift and the design of sampling policies. Collectively, these results and insights reveal a rich landscape of choices and nuances in this seemingly simple problem.

## 2. Preliminaries

### 2.1 MDPs and POMDPs

**Markov Decision Processes (MDPs)** We consider  $H$ -step finite-horizon MDPs, defined by the state space  $\mathcal{S}$ , action space  $\mathcal{A}$ , reward function  $R : \mathcal{S} \rightarrow [0, R_{\max}]$ , transition function  $P : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$ , initial state distribution  $\rho_0 \in \Delta(\mathcal{S})$ ; here we assume  $\mathcal{S}, \mathcal{A}$  are finite for convenience, and  $\Delta(\cdot)$  is the probability simplex. We adopt the convention of layered state space that allows for time-homogeneous notation for time-inhomogeneous quantities: that is, let  $\mathcal{S} = \bigcup_{0 \leq t \leq H} \mathcal{S}_t$ , where  $\rho_0$  is supported on  $\mathcal{S}_0$ .  $P(s'|s, a)$  is always 0 unless  $s \in \mathcal{S}_t$  and  $s' \in \mathcal{S}_{t+1}$ , thus any state that may appear as  $s_t$  always belongs to  $\mathcal{S}_t$ . Any policy  $\pi : \mathcal{S} \rightarrow \Delta(\mathcal{A})$  (note that this captures time-inhomogeneous policies) induces a distribution over the trajectory (or episode)  $s_0, a_0, r_0, \dots, s_{H-1}, a_{H-1}, r_{H-1}, s_H$  by the following generative process:  $s_0 \sim \rho_0, \forall t \geq 0, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim P(\cdot|s_t, a_t), r_t = R(s_{t+1})$ . We use  $\mathbb{P}_{P\pi}[\cdot]$  and  $\mathbb{E}_{P\pi}[\cdot]$  to denote this distribution and the expectation w.r.t. it.<sup>1</sup>

A standard objective that measures the performance of a policy  $\pi$  is the expected return,  $J(\pi) := \mathbb{E}_{P\pi}[\sum_{t \geq 0} r_t]$ . Let  $V_{\max} = HR_{\max}$  denote the range of the cumulative rewards. As a central concept in RL, a (Q)-value function is defined as  $Q_P^\pi(s, a) = \mathbb{E}_{P\pi}[\sum_{t' \geq t} r_{t'} | s_t = s, a_t = a]$  for  $s \in \mathcal{S}_t$ .

**Partially Observable MDPs (POMDPs)** A POMDP  $\Gamma$  is specified by an underlying MDP plus an emission process,  $E : \mathcal{S} \rightarrow \Delta(\mathcal{O})$ , which generates observation  $o_t \in \Delta(\mathcal{O})$  based on a latent state  $s_t \in \mathcal{S}$  as  $o_t \sim E(\cdot|s_t)$ ; similar to before we assume  $\mathcal{O}$  is layered, i.e.,  $o_t$  is always supported on  $\mathcal{O}_t$ . An episode in a POMDP is generated similarly to the MDP:  $s_0 \sim \rho_0$ , and at any time step  $t$ , an observation is generated as  $o_t \sim E(\cdot|s_t)$ , the agent takes action  $a_t$  that only depends on the observable history  $o_{0:t} := \{o_0, \dots, o_t\}$  and  $a_{0:t-1}$ . Then, a latent transition  $s_{t+1} \sim P(\cdot|s_t, a_t)$  occurs and a reward  $r_t$  is generated, and so on and so forth. We assume that the information of reward  $r_t$  is always encoded in  $o_{t+1}$ , so with a slight abuse of notation we write  $r_t = R(s_{t+1}) = R(o_{t+1})$ . POMDPs are often used to model processes where the observations violate the Markov property. That is, we only observe  $o_t$  in the real system, and introduce  $s_t$  to explain the dynamics and evolution of  $o_t$ . In this case,  $s_t$  are latent states that are not observed in the real-system traces.

**Belief States and History-based MDP** A key concept in POMDPs is the *belief state*,  $\mathbf{b}^*(s|\tau) := \mathbb{P}_\Gamma[s_t = s | \tau_t = \tau]$ , where  $\tau_t = (o_{0:t}, a_{0:t-1})$  is an *observable history*. It is useful to think of the evolution of the observable variables of a POMDP as a *history-based MDP*. That is, let the  $t$ -step history  $\tau_t$  be the state, and upon action  $a_t$ , the reward  $r_t$  and next-state  $\tau_{t+1}$  are generated as

$$s_t \sim \mathbf{b}^*(\cdot|\tau_t), s_{t+1} \sim P(\cdot|s_t, a_t), r_t = R(s_{t+1}), o_{t+1} \sim E(\cdot|s_{t+1}), \tau_{t+1} = (o_{0:t+1}, a_{0:t}).$$

We use  $M_\Gamma(o_{t+1}|\tau_t, a_t)$  to denote the conditional distribution and the induced MDP dynamics. (Note that since reward  $r_t$  is encoded in  $o_{t+1}$ , it can also be determined from  $\tau_{t+1}$  and thus is consistent with our MDP formulation.) This MDP naturally fits the layered

---

1. The distribution also depends on  $\rho_0$  and the reward function  $R$ , but the different models we will consider often only differ in the transition, so we use the subscript to emphasize the dependence on transition.

convention, where  $\mathcal{H}_t$ , the space of  $\tau_t$ , is the  $t$ -th step state space. This way, any history-dependent policy is simply a Markov policy in the history-based MDP,  $\pi : \mathcal{H} \rightarrow \Delta(\mathcal{A})$ . Concepts such as value functions for a POMDP can be immediately defined through its history-based MDP, that is, when we mention the Q-function in a POMDP such as  $Q_\Gamma^\pi$ , what we mean is  $Q_\Gamma^\pi = Q_{M_\Gamma}^\pi$ .

**Additional Notation** For two distributions  $p, q \in \Delta(\mathcal{X})$ , define their Total-Variation (TV) distance as  $D_{\text{TV}}(p, q) := \sum_{x \in \mathcal{X}} |p(x) - q(x)|/2$ , and let  $\|p/q\|_\infty := \max_{x \in \mathcal{X}} p(x)/q(x)$ .

## 2.2 Model Selection of Belief-State Approximations

As mentioned in the introduction, we are interested in complex simulators where, when modeled as a POMDP, the latent-state and the observation spaces  $\mathcal{S}$  and  $\mathcal{O}$  are potentially very large, and the latent transition and the emission process  $P$  and  $E$  are complex black-boxes, to which we only have sampling access. While the notion of belief state,  $\mathbf{b}^*$ , is conceptually and information-theoretically well-defined, it is not easy to access them in a computationally efficient manner, and methods from ABC, SBI, and particle filtering may be used to approximate the said belief state (Cranmer et al., 2020). Since these methods often require domain-specific design choices and heuristics, the model-selection problem naturally arises: given a candidate set of belief-state approximations  $\mathcal{B} = \{\mathbf{b}^{(i)}\}_{i=1}^m$  with  $\mathbf{b}^{(i)} : \mathcal{H} \rightarrow \Delta(\mathcal{S})$ , we are interested in selecting the best approximation by interacting with the simulator. Throughout the paper, we will assume *realizability* as a simplification:

**Assumption 1** (Realizability).  $\mathbf{b}^* \in \mathcal{B}$ .

## 3. Selection of $s_t|\tau_t$ (“Latent State-based Selection”)

The problem of selecting/learning the posterior distribution in a computationally-efficient manner is closely related to Simulation-based Inference (SBI). As a standard approach in SBI, we can generate trajectories with latent states in the form of  $(s_{0:H}, o_{0:H}, a_{0:H-1})$  using some *behavior policy*  $\pi_b : \mathcal{H} \rightarrow \Delta(\mathcal{A})$ , and obtain  $\tau_t = (o_{0:t}, a_{0:t-1})$  and  $s_t$  pairs for any  $0 \leq t \leq H$ . The joint distribution of  $(\tau_t, s_t)$  generated in this way satisfies that

$$s_t \sim \mathbf{b}^*(\cdot|\tau_t),$$

and we write  $\tau_t \sim \Gamma^{\pi_b}$  to denote that  $\tau_t$  is generated from a trajectory induced by policy  $\pi_b$  in POMDP  $\Gamma$ . On the other hand, for any candidate  $\mathbf{b} \in \mathcal{B}$ , we can also generate

$$\tilde{s}_t \sim \mathbf{b}(\cdot|\tau_t)$$

for each  $\tau_t$  in the above dataset. Then, testing whether  $\mathbf{b} = \mathbf{b}^*$  can be reduced to the problem of *conditional 2-sample test* based on the samples  $\{(\tau_t, s_t, \tilde{s}_t)\}$ , that is, to tell whether  $s_t|\tau_t$  and  $\tilde{s}_t|\tau_t$  are identically distributed.<sup>2</sup> We call this approach LATENT STATE-BASED SELECTION to distinguish it from alternative approaches we will consider later.

---

2. Strictly speaking, 2-sample test is different from and arguably harder than the selection problem, since we can leverage the realizability assumption in selection.

**Reduction to Joint 2-Sample Tests** A naïve approach is to reduce the *conditional* test to a *joint* test: we can simply test if  $(\tau_t, s_t)$  is identically distributed as  $(\tau_t, \tilde{s}_t)$ . If the joints are identical, it implies that the conditionals are also identical on the supported  $\tau_t$ . Unfortunately, this approach comes with significant practical hurdles: 2-sample tests often involve some kind of discriminator class  $\mathcal{F}$  that need to be carefully designed (Gretton et al., 2012), which in this case operates on  $\mathcal{H} \times \mathcal{S}$ . However, given that a history  $\tau \in \mathcal{H}$  is a combinatorial object of variable length, designing effective discriminators can be practically challenging. This begs the question:

*Can we design algorithms that do not rely on discriminators over the  $\mathcal{H}$  space?*

### 3.1 Selection of $Y|X$ Conditionals with $Y$ -only Discriminators

We now provide a solution to the general problem of selecting from conditional distributions in the form of  $P(Y|X)$ , in a way that only requires discriminator classes operating on  $Y$ , avoiding the demanding task of feature engineering or neural architecture design over  $X$  which are often complex combinatorial objects (such as histories) in our settings of interest.

**General Problem Formulation** We are given  $n$  i.i.d.  $(X, Y)$  pairs,  $\{(X_j, Y_j)\}$ , sampled from a real joint distribution  $(X, Y) \sim P^*$ , and the task is to select from  $m$  candidate conditionals  $P_i(y|x)$  where  $P_{i^*}(y|x) = P^*(y|x)$  for some  $i^* \in [m]$ . Computationally, we assume we can efficiently sample from  $P_i(\cdot|x)$  for any given  $x$ , but we can only sample joints from  $P^*$ .

Inspired by Scheffé tournament (Devroye and Lugosi, 2001), we first consider the case of  $m = 2$  and later extend to general  $m$  via a tournament procedure of pairwise comparison.

**Pairwise Comparison between 2 Candidates** When  $m = 2$ , we propose the following procedure, which requires a discriminator class  $\mathcal{F} : \mathcal{Y} \rightarrow \{0, 1\}$  to serve as *classifiers*: for each  $X_j$  in the data sampled from  $P^*$ ,

1. Sample  $N$  i.i.d. data points from  $P_k(\cdot|X_j)$  for  $k = 1, 2$ .
2. Use the above  $2N$  data points to train a classifier  $\hat{f}_j \in \mathcal{F}$  that predicts whether  $Y$  is sampled from  $P_1(\cdot|X_j)$  or  $P_2(\cdot|X_j)$ . For theoretical analyses and presentation ease, we assume ERM on 0/1 loss, and adopt the convention (which is w.l.o.g.) that  $P_{i^*}$  gets label 1.
3. Use  $\hat{f}_j$  to classify the real  $Y_j$ . Additionally sample 1 data point from each of  $P_1(\cdot|X_j)$  and  $P_2(\cdot|X_j)$ , denoted as  $Y_j^{(1)}$  and  $Y_j^{(2)}$ , and classify them with  $\hat{f}_j$  as well.

Finally, we choose between  $P_1, P_2$  based on  $\arg \min_{k \in \{1, 2\}} \left| \frac{1}{n} \sum_j \hat{f}_j(Y_j) - \frac{1}{n} \sum_j \hat{f}_j(Y_j^{(k)}) \right|$ . For  $m > 2$ , we perform the above procedure for each pair of candidate conditionals (data sampled from  $P_k$  may be reused in multiple comparisons), and let  $\hat{f}_j^{i,k}$  be the classifier

trained to distinguish between  $P_i(\cdot|X_j)$  and  $P_k(\cdot|X_j)$ . The final choice is

$$\arg \min_{i \in [m]} \max_{k \in [m], k \neq i} \left| \frac{1}{n} \sum_j \hat{f}_j^{i,k}(Y_j) - \frac{1}{n} \sum_j \hat{f}_j^{i,k}(Y_j^{(i)}) \right|. \quad (1)$$

In words, for each real data point  $X_j$ , we draw “synthetic data” from the candidate conditionals  $P_i(\cdot|X_j)$  and  $P_k(\cdot|X_j)$  to train a classifier, and apply it to a single “real” data point  $Y_j$ ; in total,  $nm(m-1)/2$  classifiers will be trained. When  $i^* \in \{i, k\}$ , the classifier learns to separate  $Y$  generated using  $P^* = P_{i^*}$  from that using the wrong conditional. Therefore, we may choose between  $P_i$  and  $P_k$  based on  $\hat{f}_j^{i,k}(Y_j)$ , which predicts whether  $Y_j$  (sampled from  $P^* = P_{i^*}$ ) is more likely to be produced by  $P_i$  or  $P_k$ . Of course, the signal from classifying a single data point  $Y_j$  is weak and contains randomness, and aggregating such signals across all  $j$  can reduce the noise and amplify the signal.

While the above idea is reasonable, it may run into issues when the data from  $P_i$  and  $P_k$  are not cleanly separable by  $\mathcal{F}$ : the algorithm minimizes an overall error rate with a mixture of correct data (from  $P_{i^*}$ ) and incorrect data, but the classifier is eventually only applied to  $Y_j$  from  $P_{i^*}$ , meaning that the ultimate loss we suffer is the False Negative Rate of  $\hat{f}_j^{i,k}$ , which is only loosely controlled by the overall error rate. In contrast, we follow the spirit of Scheffé estimators and treat the classifier  $\hat{f}_j^{i,k}$  as an approximate witness of the total-variation distance between  $P_i(\cdot|X_j)$  and  $P_k(\cdot|X_j)$ , and make the final selection via the IPM loss in Eq.(1), which leads to the relatively weak assumption in the theoretical analysis below.

**Theoretical Guarantee** We now provide the assumptions and the theoretical guarantees for this procedure. The key assumption we need is that  $\mathcal{F}$  contains nontrivial classifiers that separate  $P_i(\cdot|X_j)$  from  $P_k(\cdot|X_j)$  when  $i^* \in \{i, k\}$ , as formalized by the following assumption.

**Assumption 2** (Expressivity of  $\mathcal{F}$ ). Assume  $\mathcal{F}$  is a finite class. Define

$$\text{acc}_X^{i,i^*}(f) := 1/2 \cdot \Pr_{Y \sim P^*(\cdot|X)}[f_X^{i,i^*}(Y) = 1] + 1/2 \cdot \Pr_{Y \sim P_i(\cdot|X)}[f_X^{i,i^*}(Y) = 0].$$

For any  $i \neq i^*$ , let

$$f_{x,\star}^{i,i^*}(y) = \mathbb{I}[P_{i^*}(y|x) > P_i(y|x)] \quad (2)$$

be the Bayes-optimal classifier for distinguishing between  $P_{i^*}(\cdot|x)$  and  $P_i(\cdot|x)$ , and

$$\mathcal{E}(i, i^*) = \mathbb{E}_{X \sim P^*} \left[ \text{acc}_X^{i,i^*}(f_{X,\star}^{i,i^*}) \right] - 1/2.$$

We assume the existence of  $f_X^{i,i^*} \in \mathcal{F}$  that satisfies the following: for some  $0 < \alpha \leq 1$  that applies to all  $i$ ,

$$\mathbb{E}_{X \sim P^*} \left[ \text{acc}_X^{i,i^*}(f_X^{i,i^*}) \right] \geq 1/2 + \alpha \mathcal{E}(i, i^*).$$

$f_{X,\star}^{i,i^\star}$  (Eq.(2)) is the best possible classifier for separating  $P_{i^\star}(\cdot|X) = P^\star(\cdot|X)$  from  $P_i(\cdot|X)$  when  $i \neq i^\star$ , and we can get straightforward guarantees if we simply assume  $f_{X,\star}^{i,i^\star} \in \mathcal{F}$ . Instead, we only assume  $\mathcal{F}$  realizes “better-than-trivial” classifier  $f_X^{i,i^\star}$ , and the rest of this assumption introduces definitions to quantify what “better-than-trivial” means.

The  $\text{acc}_X^{i,i^\star}$  term is the classification accuracy, with the convention (which is w.l.o.g.) that  $P_{i^\star}$  is assigned label 1 and  $P_i$  is assigned label 0. Note that when  $\mathcal{F}$  is closed under negation ( $1-f \in \mathcal{F}, \forall f \in \mathcal{F}$ ), it is trivial to find a classifier in  $\mathcal{F}$  with 1/2 accuracy, so  $\mathcal{E}(i, i^\star)$  is a measure of how separable  $P_i$  and  $P_{i^\star}$  are; in fact,  $\mathcal{E}(i, i^\star) = 1/2 \cdot \mathbb{E}_{X \sim P^\star}[D_{\text{TV}}(P_i(\cdot|X), P_{i^\star}(\cdot|X))]$ . Given all these definitions, we require  $\mathcal{F}$  to contain a classifier  $f_X^{i,i^\star}$  whose margin is only a multiplicative fraction of  $\mathcal{E}(i, i^\star)$ , and this does not need to hold for every  $X$ , but only on average w.r.t. the marginal of  $X$ .

We are now ready to give the guarantee; the proof is deferred to Appendix A.

**Theorem 1** (Sample complexity). *Under Assumption 2, for  $\hat{i}$  identified by Eq.(1), with probability at least  $1 - \delta$ ,*

$$\mathbb{E}_{X \sim P^\star}[D_{\text{TV}}(P_{\hat{i}}(\cdot|X), P^\star(\cdot|X))] \leq \epsilon,$$

as long as

$$n = O\left(\frac{\log(m/\delta)}{\alpha^2 \epsilon^2}\right), \quad N = O\left(\frac{\log(mn|\mathcal{F}|/\delta)}{\alpha^2 \epsilon^2}\right).$$

Invoked on  $X = \tau_t$ ,  $Y = s_t$ , and  $P^\star$  is distribution under behavior policy  $\pi_b$ , we have

$$\mathbb{E}_{\tau_t \sim \Gamma^{\pi_b}}[D_{\text{TV}}(\mathbf{b}(\cdot|\tau_t), \mathbf{b}^\star(\cdot|\tau_t))] \leq \epsilon, \quad (3)$$

where  $\tau_t \sim \Gamma^{\pi_b}$  is a partial trajectory naturally simulated in  $\Gamma$  under policy  $\pi_b$  without using resetting.

**Related Works** The above procedure is closely related to and a variant of the method of Li et al. (2022), both of which can be viewed as the extension of Scheffé estimators to conditional distributions. The difference is that Li et al. assume access to the density functions of  $P_i(y|x)$ , which allows them to explicitly compute the Bayes-optimal classifier in Eq.(2). In contrast, we only allow blackbox sampling access to  $P_i(\cdot|x)$ , and approximate the above classifier using a discriminator class  $\mathcal{F}$  via training on sampled synthetic data. What we show is that  $\mathcal{F}$  does not need to realize the above Bayes-optimal classifier for every single  $x$ , and it suffices to have marginally-better-than-trivial classifiers in an average sense. Theoretically, Li et al. show that the sample complexity of the conditional selection problem should not depend on the complexity of the  $\mathcal{X}$  space (see also Bilodeau et al. (2023)), which is echoed by our motivation of not having to design discriminators over  $\mathcal{X}$ . In model-based RL, similar insights and Scheffé-style constructions have been used in learning model dynamics from IPM losses (Sun et al., 2019). The idea of discriminators to help learn or test conditional distributions are also found in the SBI literature (Lueckmann et al., 2021).

Outside belief-state approximation, our procedure may also be relevant to model selection in conditional generative models, such as prompt-based image generation. One potentially



relevant property is that our procedure has a low sample-complexity burden on  $n$ , the number of “real” data points from  $P^*$ . In belief-state approximation, both  $n$  and  $N$  can be increased by spending more computation; in tasks of learning from real datasets, however, the real data (from  $P^*$ ) can be more expensive to collect compared to the synthetic data (from  $P^i$ ), and the independence of  $n$  on the complexity of  $|\mathcal{F}|$  can be appealing.

#### 4. Selection of $o_{t+1}|\tau_t, a_t$ (“Observation-based Selection”)

We now show that the dynamical-system nature of POMDPs adds interesting twists to the problem and allows for alternative solutions. In particular, we show that choosing the right conditional of  $s_t|\tau_t$  (by reducing to the problem of selecting the conditional distribution of  $Y|X$  with  $X = \tau_t$  and  $Y = s_t$ ) is not the only way to select the belief state approximation. Instead, we can select for the right *observable model* induced by the belief state approximations.

**Observable Model** A POMDP  $\Gamma$  and an approximate belief state  $\mathbf{b}$  together defines an *observable model*  $M_{\Gamma, \mathbf{b}}$ , which is a mapping in the form of  $\mathcal{H} \times \mathcal{A} \rightarrow \Delta(\mathcal{O})$ . Using this model  $M_{\Gamma, \mathbf{b}}$ , we can efficiently sample the next observation  $o_{t+1}$  given an history  $\tau_t$  and action  $a_t$ , as described by the following sampling process:  $o_{t+1} \sim M_{\Gamma, \mathbf{b}}(\cdot|\tau_t, a_t)$  is equivalent to

$$s_t \sim \mathbf{b}(\cdot|\tau_t), s_{t+1} \sim P(\cdot|s_t, a_t), o_{t+1} \sim E(\cdot|s_{t+1}). \quad (4)$$

This model can be equivalently treated as a history-based MDP, as it describes how the next state (length- $(t+1)$  history) can be sampled from the current state-action pair (length- $t$  history and action): given  $\tau_t, a_t$ , the generative process for  $\tau_{t+1}$  is simply

$$o_{t+1} \sim M_{\Gamma, \mathbf{b}}(\cdot|\tau_t, a_t), \quad \tau_{t+1} = \tau_t \circ a_t \circ o_{t+1},$$

where  $\circ$  is concatenation. It is not hard to see that when  $\mathbf{b} = \mathbf{b}^*$ ,  $M_{\Gamma, \mathbf{b}^*}$  describes the true history-based MDP induced by POMDP  $\Gamma$ , i.e.,  $M_{\Gamma, \mathbf{b}^*} = M_{\Gamma}$ . Given that most key RL concepts in POMDPs, such as value functions and optimal policies, can be defined through the induced observable model (see Section 2.1), a natural idea is to apply the conditional selection procedure in Section 3 but with the following setup:

- $X = (\tau_t, a_t)$ ,  $Y = o_{t+1}$ .
- $P^*$  correspond to generating  $X, Y$  pairs by sampling trajectories using some behavior policy  $\pi_b : \mathcal{H} \rightarrow \Delta(\mathcal{A})$ .
- Each candidate  $\mathbf{b}^{(i)}$  induces a conditional  $P_i(\cdot|X) = M_{\Gamma, \mathbf{b}^{(i)}}(\cdot|\tau_t, a_t)$ .

Then under Assumption 2, we can directly invoke Theorem 1 for an expected TV guarantee. For example:

**Corollary 2.** *Bind  $X = (\tau_t, a_t)$  and  $Y = o_{t+1}$  and define  $P^*$ ,  $\{P_i\}$  as described above. Under the conditions of Theorem 1, with probability at least  $1 - \delta$ , we will select a belief state approximation  $\mathbf{b}$ , such that (note that  $M_{\Gamma} = M_{\Gamma, \mathbf{b}^*}$ )*

$$\mathbb{E}_{(\tau_t, a_t) \sim \Gamma^{\pi_b}} [D_{\text{TV}}(M_{\Gamma, \mathbf{b}}(\cdot|\tau_t, a_t), M_{\Gamma}(\cdot|\tau_t, a_t))] \leq \epsilon. \quad (5)$$



Furthermore, such a guarantee are immediately implied from (and thus weaker than) that for LATENT STATE-BASED SELECTION:

**Proposition 3.** *Fixing any  $\tau_t, a_t$ , we have*

$$D_{\text{TV}}(M_{\Gamma, \mathbf{b}}(\cdot|\tau_t, a_t), M_{\Gamma}(\cdot|\tau_t, a_t)) \leq D_{\text{TV}}(\mathbf{b}(\cdot|\tau_t), \mathbf{b}^*(\cdot|\tau_t)).$$

*Therefore, Eq.(3) immediately implies Eq.(5).*

*Proof.* Conditioned  $s_t$ , the process of generating  $o_{t+1}$  is independent of whether  $s_t$  is generated from  $\mathbf{b}^*$  or  $\mathbf{b}$ , so the claim is a direct consequence of the data processing inequality.  $\square$

If the final goal is to compute objects that solely depend on the observable model  $M_{\Gamma}$ , such as value functions or optimal policies, it seems that OBSERVATION-BASED SELECTION is an equally legitimate solution. In fact, OBSERVATION-BASED SELECTION has an additional advantage of being more robust to misspecification: if the latent state  $s_t$  includes dummy variables that do not affect the observable dynamics, OBSERVATION-BASED SELECTION can effectively ignore the errors of  $\mathbf{b}$  on these inconsequential variables and thus require a weaker version of realizability, while LATENT STATE-BASED SELECTION still treats such  $\mathbf{b}$  as incorrect and insists on choosing the groundtruth  $\mathbf{b}^*$ .

This OBSERVATION-BASED SELECTION approach reflects a prevailing theme in RL research on POMDPs, namely *behavioral equivalence*, that the observable behavior of a POMDP is what ultimately matters, and latent states are *ungrounded* objects and only a convenient way to help express the observable behavior (e.g., the latent state transition  $P$  and emission  $E$  are a convenient way to parameterize the observable dynamics  $M_{\Gamma}$ ). This philosophy is most pronounced in the research of Predictive State Representations (Littman and Sutton, 2002; Singh et al., 2004) and minimal information state (Subramanian et al., 2022), and is also manifested in recent statistical results for learning in POMDPs (Zhang and Jiang, 2025) (see also Section 5.3). As we will see next, however, OBSERVATION-BASED SELECTION can suffer from surprising degeneracy in downstream use cases when LATENT STATE-BASED SELECTION is well-behaved, which seems to contradict the spirit of behavioral equivalence.<sup>3</sup>

## 5. Roll-Out Guarantees and Temporal Consistency

In this section, we investigate how the errors in  $\mathbf{b}$  (measured in the forms of Eq.(3) or (5)) can affect the downstream RL tasks. We consider the most basic yet representative use of state resetting enabled by an approximate  $\mathbf{b}$ :

**Roll-out:** Sample trajectories to obtain a Monte-Carlo estimate of  $Q_{\Gamma}^{\pi}(\tau_t, a_t)$  for a given target policy  $\pi$ .

This procedure is useful for debugging and simulator selection (Sajed et al., 2018; Liu et al., 2025), enables Monte-Carlo control (Sutton and Barto, 2018), and forms the basis

---

3. The paradox is resolved by realizing that we rely on  $P$  and  $E$  for efficient sampling which ground the latent states.

of Monte-Carlo Tree Search (Kocsis and Szepesvári, 2006; Silver and Veness, 2010; Browne et al., 2012). Perhaps surprisingly, despite the simplicity of the task, there are many nuances to the question. LATENT STATE-BASED SELECTION and OBSERVATION-BASED SELECTION interact in subtle but consequential ways with the roll-out method, and OBSERVATION-BASED SELECTION can have degenerate behaviors when used with the most natural roll-out procedure.

### 5.1 Single-Reset Roll-Out

We now consider what error guarantees can be obtained for estimating  $Q_{\Gamma}^{\pi}(\tau_t, a_t)$  via Monte-Carlo roll-outs using an approximate belief state  $\mathbf{b} \approx \mathbf{b}^*$ , e.g., one with guarantees established in Theorem 1. In particular, the obvious (and *seemingly* unique) roll-out procedure, which we call “SINGLE-RESET Roll-out” (whose meaning will be clear shortly), is:

**Single-Reset Roll-out.** Input:  $\tau_t, a_t, \pi$ .

1. Sample  $s_t \sim \mathbf{b}(\cdot|\tau_t)$ .
2. Take given action  $a_t$ , and generate  $s_{t+1} \sim P(\cdot|s_t, a_t)$ ,  $o_{t+1} \sim E(\cdot|s_{t+1})$ .
3. Repeat Step 2 for subsequent time steps by taking actions according to  $\pi$ , and collect  $\sum_{t'=t+1}^H R(o_{t'})$  as a Monte-Carlo return.

Since the error due to finite roll-outs can be easily handled by Hoeffding’s inequality, we will focus on the expected value (i.e., assuming infinitely many roll-outs) obtained through the above procedure, denoted as  $Q_{1\text{-RESET}(\Gamma, \mathbf{b})}^{\pi}(\tau_t, a_t)$ , and analyze its error relative to the groundtruth  $Q_{\Gamma}^{\pi}(\tau_t, a_t)$ . Concretely, the guarantee in the form of Eq.(3), obtained via LATENT STATE-BASED SELECTION, immediately implies the proposition below. Since our guarantee for  $\mathbf{b}$  in Eq.(3) is not pointwise, it should not be surprising that the error bound for  $Q_{1\text{-RESET}(\Gamma, \mathbf{b})}^{\pi}$  is also given in an average sense w.r.t. the distribution of trajectories under  $\pi_b$ , since that is what we use to train the classifiers and select for  $s_t|\tau_t$ .

**Proposition 4.** *If Eq.(3) (guarantee for LATENT STATE-BASED SELECTION) holds for some  $t$ , then*

$$\mathbb{E}_{(\tau_t, a_t) \sim \Gamma^{\pi_b}} [|Q_{1\text{-RESET}(\Gamma, \mathbf{b})}^{\pi}(\tau_t, a_t) - Q_{\Gamma}^{\pi}(\tau_t, a_t)|] \leq \epsilon \cdot V_{\max}.$$

*Proof.* The procedure for rolling out using  $\mathbf{b}$  vs.  $\mathbf{b}^*$  is identical after  $s_t$  is sampled, so  $\mathbb{E}[\sum_{t'=t+1}^H R(o_{t'})|\tau_t, a_t, s_t; \pi]$  is the same for both  $\mathbf{b}$  and  $\mathbf{b}^*$ , where the expectation is w.r.t. the randomness of the SINGLE-RESET procedure. The final result is just taking its expectation w.r.t.  $s_t \sim \mathbf{b}(\cdot|\tau_t)$  vs.  $s_t \sim \mathbf{b}^*(\cdot|\tau_t)$ , so

$$|Q_{1\text{-RESET}(\Gamma, \mathbf{b})}^{\pi}(\tau_t, a_t) - Q_{\Gamma}^{\pi}(\tau_t, a_t)| \leq D_{\text{TV}}(\mathbf{b}(\cdot|\tau_t), \mathbf{b}^*(\cdot|\tau_t)) \cdot V_{\max}.$$

Plugging this into  $\mathbb{E}_{(\tau_t, a_t) \sim \Gamma^{\pi_b}} [\cdot]$  completes the proof.  $\square$

As a natural follow-up question, can we obtain similar error bounds from Eq.(5), which is obtained from OBSERVATION-BASED SELECTION?

## 5.2 Repeated-Reset Roll-out

Recall that OBSERVATION-BASED SELECTION enjoys the guarantee in Eq.(5):

$$\mathbb{E}_{(\tau_t, a_t) \sim \Gamma^{\pi_b}} [D_{\text{TV}}(M_{\Gamma, \mathbf{b}}(\cdot | \tau_t, a_t), M_{\Gamma}(\cdot | \tau_t, a_t))] \leq \epsilon.$$

It is tempting to provide guarantees for the expected roll-out value by directly reducing to existing MDP analysis, using the following argument: note that  $M_{\Gamma, \mathbf{b}}$  and  $M_{\Gamma, \mathbf{b}^*}$  are two history-based MDPs. When we treat history as state, Eq.(5) is similar to the kind of guarantees obtained from MLE in MDPs,<sup>4</sup> which immediately leads to a policy-evaluation guarantee:

**Theorem 5.** *Suppose two MDPs only differ in the transition dynamics,  $P$  vs.  $P'$ . In addition, for every  $t$ , assume*

$$\mathbb{E}_{(s_t, a_t) \sim P^{\pi_b}} [D_{\text{TV}}(P'(\cdot | s_t, a_t), P(\cdot | s_t, a_t))] \leq \epsilon. \quad (6)$$

*When the roll-out policy  $\pi = \pi_b$ , it holds that for every  $t$ ,*

$$\mathbb{E}_{(s_t, a_t) \sim P^{\pi_b}} [|Q_{P'}^{\pi}(s_t, a_t) - Q_P^{\pi}(s_t, a_t)|] \leq \epsilon(H - t)V_{\max}.$$

This result is standard in model-based RL, and we provide its proof in Appendix B for completeness. It is also possible to extend the result to other roll-out policies  $\pi \neq \pi_b$  by paying for a coverage coefficient, which we will discuss in Section 6.1 but not consider in this section. The exact match between Eq.(6) and Eq.(5) immediately implies that:

**Corollary 6.** *If Eq.(5) holds for all  $t$ , then for  $\pi = \pi_b$  we have (note that  $Q_{\Gamma}^{\pi} = Q_{M_{\Gamma, \mathbf{b}^*}}^{\pi}$ ):*

$$\mathbb{E}_{(\tau_t, a_t) \sim \Gamma^{\pi_b}} [|Q_{M_{\Gamma, \mathbf{b}}}^{\pi}(\tau_t, a_t) - Q_{\Gamma}^{\pi}(\tau_t, a_t)|] \leq \epsilon(H - t)V_{\max}.$$

While the guarantee here has an additional horizon factor  $(H - t)$ , it is nevertheless a solid guarantee. The final piece of the puzzle is the observation that SINGLE-RESET *seems* to be the only way to sample from  $\Gamma$  using  $\mathbf{b}$ , so it *probably* coincides with  $M_{\Gamma, \mathbf{b}}$ , i.e.,  $Q_{1\text{-RESET}(\Gamma, \mathbf{b})}^{\pi} = Q_{M_{\Gamma, \mathbf{b}}}^{\pi}$  for any  $\pi$ .

This intuitive statement, however, is in sharp conflict with the following example:

**Example 1** ( $Q_{1\text{-RESET}(\Gamma, \mathbf{b})}^{\pi_b}$  **cannot** enjoy the guarantee of Corollary 6). *Consider an arbitrary POMDP  $\Gamma$ , except that the emission  $E(\cdot | \cdot)$  is state-independent at some time step  $t_0$ . Every candidate  $\mathbf{b}$  predicts the correct belief state  $\mathbf{b}^*(\cdot | \tau_t)$  for any  $t \neq t_0$ , but can predict arbitrary distributions for  $t = t_0$ . In this case, the observable behavior (i.e., the induced law of  $M_{\Gamma, \mathbf{b}}$ ) of  $\mathbf{b}$  is indistinguishable from that of  $\mathbf{b}^*$  at any time step, so any  $\mathbf{b}$  could be selected. However, a SINGLE-RESET roll-out starting at  $t_0 - 1$  will generally be incorrect since the arbitrarily generated  $s_{t_0-1}$  will have a lingering effect, i.e., producing an incorrect distribution of  $s_{t_0}$  and subsequent latent states.*

We are facing a paradox: on one hand, Example 1 shows that OBSERVATION-BASED SELECTION cannot enjoy the guarantee in Corollary 6 for SINGLE-RESET roll-out; on the other hand, the observable model  $M_{\Gamma, \mathbf{b}}$ , a history-based MDP determined by  $\Gamma$  and  $\mathbf{b}$ , does enjoy

---

4. The guarantee of MLE has a square on the TV distance, i.e.,  $\mathbb{E}[D_{\text{TV}}(\cdot)^2] \leq \dots$

Real trace:	X O O   O X O O O O X O O X
SINGLE-RESET:	X O O   X O O O O X O O X O
REPEATED-RESET:	X O O   X X X X X X X X X X

Figure 1: Toy example for illustrating the difference between SINGLE-RESET and REPEATED-RESET. In this binary-observation, action-less system, “X” represents the occurrence of an event. Every time an event happens (“X”), the system samples the interval till next event from some distribution. The history of interest is “X O O”, and the first row shows the real trajectory.  $\mathbf{b}$  always sets the latent state to be 0, i.e., predicts that next event will occur immediately.

a standard guarantee for its induced Q-value. Given that SINGLE-RESET seems to be the only reasonable way to roll-out trajectories using  $\Gamma$  and  $\mathbf{b}$ , it is reasonable to believe that such roll-outs correspond to the Q-value of  $M_{\Gamma, \mathbf{b}}$ . So what gives?

**Fact 1.**  $Q_{1\text{-RESET}(\Gamma, \mathbf{b})}^\pi$  is *not* equivalent to  $Q_{M_{\Gamma, \mathbf{b}}}^\pi$ .

The reason why these two objects are different should be made clear by the following procedure that actually produces roll-outs according to  $M_{\Gamma, \mathbf{b}}$ :

**Repeated-Reset Roll-out.**      Input:  $\tau_t, a_t, \pi$ .

1. Sample  $s_t \sim \mathbf{b}(\cdot|\tau_t)$ .
2. Take given action  $a_t$ , and generate  $s_{t+1} \sim P(\cdot|s_t, a_t)$ ,  $o_{t+1} \sim E(\cdot|s_{t+1})$ .
3. **Replace  $s_{t+1}$  with a fresh sample from  $\mathbf{b}(\cdot|\tau_{t+1})$**  where  $\tau_{t+1} = \tau_t \circ a_t \circ o_{t+1}$ .
4. Repeat Steps 2 and 3 for subsequent time steps by taking actions according to  $\pi$ , and collect  $\sum_{t'=t+1}^H R(o_{t'})$  as a Monte-Carlo return.

SINGLE-RESET and REPEATED-RESET are equivalent if  $\mathbf{b} = \mathbf{b}^*$  but are generally different; see Figure 1 for an example. We conclude this section with the following remarks that reconcile the earlier paradox:

- The main difference between SINGLE-RESET and REPEATED-RESET is that, at any time  $t$ , the observable trajectory  $\tau_t$  is a sufficient statistics for simulating the rest of the trajectory in REPEATED-RESET. For SINGLE-RESET, however, the sufficient statistic is  $(\tau_t, s_t)$ . Therefore, when we only have OBSERVATION-BASED SELECTION guarantee but not that of LATENT STATE-BASED SELECTION, SINGLE-RESET is only guaranteed to produce the correct  $o_{t+1}$ , but not future observations due to the lingering effect of possibly wrong  $s_t$ .
- OBSERVATION-BASED SELECTION can still enjoy the guarantee in Corollary 2 via REPEATED-RESET roll-out, but it suffers from an additional horizon factor compared to Proposition 4 due to the repeated application of the inaccurate  $\mathbf{b}$  (c.f. Figure 1). On a related note, Proposition 4 only requires Eq.(3) to hold for the  $t$  that is the subscript of the  $\tau_t$  we start roll-out from, but Corollary 2 requires Eq.(5) to hold for all  $t' \geq t$ . Moreover, when the roll-out policy  $\pi \neq \pi_b$ , Corollary 6 needs to pay for an additional coverage coefficient (see Section 6.1) while Proposition 4 does not. Therefore, while LATENT STATE-BASED SELECTION + REPEATED-RESET can also enjoy Corollary 2 via Proposition 3, it is in-

Table 1: Relationship between the selection methods and the roll-out methods.

	LATENT STATE-BASED SELECTION	OBSERVATION-BASED SELECTION
Agnostic to redundant latent variables	✗	✓ (see the end of Section 4)
SINGLE-RESET	✓ Proposition 4	✗ Example 1
REPEATED-RESET	✓ Corollary 2 (worse than Proposition 4)	

ferior to LATENT STATE-BASED SELECTION + SINGLE-RESET in both error propagation and computational efficiency.

- As another possible misconception, it may be tempting to think that the observable dynamics of SINGLE-RESET is  $M_{\Gamma, \mathbf{b}}$  at time step  $t$  and  $M_{\Gamma}$  for subsequent time steps, since all later simulations do not involve the use of the inaccurate  $\mathbf{b}$  and hence should be “correct”. This is not true due to, again, the lingering effect of wrong  $s_t$  distribution. (In fact, if this held, Corollary 2 would hold for  $Q_{1-\text{RESET}(\Gamma, \mathbf{b})}^{\pi}$  without the horizon factor.) That is, although all subsequent simulations seem correct, the induced conditional law of the observables,  $o_{t'+1}|\tau_{t'}, a_{t'}$  for  $t' > t$  is *not* the same as the true  $M_{\Gamma}$ .
- As potential future directions, it will be interesting to consider if the inconsistency between SINGLE-RESET and REPEATED-RESET can be turned into a method for selecting belief-state approximations, and whether the insights can be used to reduce the error accumulation of OBSERVATION-BASED SELECTION.

### 5.3 Case Study: Simulator Selection from Real-System Traces

The previous sections may leave the impression that LATENT STATE-BASED SELECTION is more superior to OBSERVATION-BASED SELECTION other than a minor disadvantage regarding redundant latent variables. Below we study a motivating scenario mentioned earlier, where it is beneficial to integrate both LATENT STATE-BASED SELECTION and OBSERVATION-BASED SELECTION into the solution and exploit the strength of each.

**Problem Setup.** Consider the problem of learning from real-system data. Let  $\Gamma_{\star}$  be a real system, from which we draw observable trajectories with policy  $\pi_b$ . The goal is to use these data trajectories to select among candidate simulators  $\{\Gamma_k\}_{k=1}^K$  that best matches the dynamics of the real system, and as before we assume realizability i.e.,  $\Gamma_{\star} \in \{\Gamma_k\}$ . In a way, this is essentially a model estimation problem in POMDPs, and the standard approach (as mentioned earlier in Section 4) is MLE (Liu et al., 2022):

$$\arg \min_k \sum_j \log M_{\Gamma_k}(o_{t+1}^{(j)} | \tau_t^{(j)}, a_t^{(j)}),$$

where all variables with  $(j)$  subscript come from the  $j$ -th data trajectory.

Unfortunately, this solution is not directly applicable to our setting, since we do not assume probability mass/density access to  $P$  or  $E$  in any given simulator  $\Gamma$  and thus cannot compute  $M_{\Gamma_k}(\cdot|\cdot)$ . Instead, we can directly leverage our OBSERVATION-BASED SELECTION (Section 4):

while it is initially designed to select  $M_\Gamma$  from  $\{M_{\Gamma, \mathbf{b}} : \mathbf{b} \in \mathcal{B}\}$ , the procedure and analyses immediately apply to the problem here where we select  $M_{\Gamma_\star}$  from  $\{M_{\Gamma_k}\}$ .

But that brings a further problem: OBSERVATION-BASED SELECTION requires efficient sampling access to the candidate conditionals, which is  $M_\Gamma(o_{t+1}|\tau_t, a_t)$  for  $\Gamma \in \{\Gamma_k\}$  here. That requires having sampling access to the belief state in  $\Gamma$  (Eq.(4)), which is not available. However, that is exactly the problem we have been dealing with so far! So let's assume that we are given  $\mathcal{B}$ ,<sup>5</sup> such that for each  $\Gamma \in \{\Gamma_k\}$ , the true belief state  $\mathbf{b}^\star \in \mathcal{B}$ .<sup>6</sup> This leads to a two-stage procedure: in **Stage 1**, for each  $\Gamma$ , we select  $\mathbf{b}_\Gamma \in \mathcal{B}$  as its belief-state approximation; in **Stage 2**, we use OBSERVATION-BASED SELECTION to select an observable model from  $\{(\Gamma_k, \mathbf{b}_{\Gamma_k})\}$ .

**Solution Solely based on Observation-based Selection** Ultimately, we need to select a  $(\Gamma, \mathbf{b})$  pair from  $\{\Gamma_k\} \times \mathcal{B}$ . Given that observable trajectories from  $\Gamma_\star$  force the use of OBSERVATION-BASED SELECTION in the second stage, a natural simplification is to lump the first stage into the second and solve both simultaneously with one unified OBSERVATION-BASED SELECTION instance. That is, we select

$$M_{\Gamma_\star} \quad \text{from} \quad \{M_{\Gamma, \mathbf{b}} : \Gamma \in \{\Gamma_k\}, \mathbf{b} \in \mathcal{B}\}.$$

Invoking the analyses in Corollaries 2 and 6, we immediately have:<sup>7</sup> for the selected  $(\Gamma, \mathbf{b})$  pair,

$$\begin{aligned} \mathbb{E}_{(\tau_t, a_t) \sim \Gamma_\star^{\pi_b}} [D_{\text{TV}}(M_{\Gamma, \mathbf{b}}(\cdot|\tau_t, a_t), M_{\Gamma_\star}(\cdot|\tau_t, a_t))] &\leq \epsilon_0, \\ \mathbb{E}_{(\tau_t, a_t) \sim \Gamma_\star^{\pi_b}} [|Q_{M_{\Gamma, \mathbf{b}}}^{\pi_b}(\tau_t, a_t) - Q_{\Gamma_\star}^{\pi_b}(\tau_t, a_t)|] &\leq \epsilon_0 H V_{\max}, \end{aligned} \quad (7)$$

where  $\epsilon_0$  can be determined by the number of data trajectories from  $\Gamma_\star$  through the sample-complexity statement in Theorem 1.<sup>8</sup>

The problem is, if we want to roll-out trajectories using the selected  $(\Gamma, \mathbf{b})$  to approximate  $Q_{\Gamma_\star}^{\pi_b}$ , the only valid approach is REPEATED-RESET (Eq.(7)), and SINGLE-RESET will not enjoy any guarantee given Example 1. However, REPEATED-RESET is computationally costly especially when sampling from  $\mathbf{b}$  has a nontrivial cost, and in practice REPEATED-RESET is often a bad idea given repeated injection of the error of  $\mathbf{b} \neq \mathbf{b}^\star$  (Figure 1). This begs the question: can we enjoy a guarantee similar to Eq.(7) while rolling out from the selected  $(\Gamma, \mathbf{b})$  using SINGLE-RESET?

**Two-Stage Solution** We now show that the natural two-stage solution achieves the goal if we use LATENT STATE-BASED SELECTION in the first stage (selecting  $\mathbf{b}$  for  $\Gamma$ ). The analysis turns out to be somewhat nontrivial, which we provide below:

- 
- 5. The candidate set  $\mathcal{B}$  can be designed for each  $\Gamma \in \{\Gamma_k\}$  separately, but we assume  $\mathcal{B}$  is the same across candidate simulators for ease of presentation.
  - 6. We still use  $\mathbf{b}^\star$  to refer to the true belief state of the simulator  $\Gamma$  under consideration. Note that we do not need to refer to the belief state of the real system  $\Gamma_\star$  and hence does not give it a notation.
  - 7. We relax  $(H - t)$  in Corollary 6 to  $H$  here for readability.
  - 8. Concretely, to achieve  $\epsilon_0$  error, the number of trajectories needed is  $O(\log(mK/\delta)/\alpha^2 \epsilon_0^2)$ , where  $m$  and  $K$  are the sizes of  $\mathcal{B}$  and  $\{\Gamma_k\}_{k=1}^K$ , respectively.

**Theorem 7.** Assume that the selected  $(\Gamma, \mathbf{b})$  satisfies

$$\mathbb{E}_{\tau_t \sim \Gamma^{\pi_b}} [D_{\text{TV}}(\mathbf{b}(\cdot|\tau_t), \mathbf{b}^*(\cdot|\tau_t))] \leq \epsilon_1. \quad (8)$$

$$\mathbb{E}_{(\tau_t, a_t) \sim \Gamma_\star^{\pi_b}} [D_{\text{TV}}(M_{\Gamma, \mathbf{b}}(\cdot|\tau_t, a_t), M_{\Gamma_\star}(\cdot|\tau_t, a_t))] \leq \epsilon_0. \quad (9)$$

Then

$$\mathbb{E}_{(\tau_t, a_t) \sim \Gamma_\star^{\pi_b}} [|Q_{1\text{-RESET}(\Gamma, \mathbf{b})}^{\pi_b}(\tau_t, a_t) - Q_{\Gamma_\star}^{\pi_b}(\tau_t, a_t)|] \leq (2\epsilon_0 + 3\epsilon_1)HV_{\max}.$$

The conditions of the theorems are the guarantees of LATENT STATE-BASED SELECTION in Stage 1 (Eq.(3)) and OBSERVATION-BASED SELECTION in Stage 2 (Eq.(5)) when  $\Gamma_\star$  is the groundtruth).<sup>9</sup> The final guarantee resembles Eq.(7), except that it permits the use of SINGLE-RESET roll-out. The error bound is slightly worse than Eq.(7) by a multiplicative constant and an additional dependence on  $\epsilon_1$ . However, note that  $\epsilon_0$  is determined by the amount of real-system data which often is fixed, while  $\epsilon_1$  is determined by the amount of data sampled from each simulator  $\Gamma$ . So overall the guarantee is still comparable to Eq.(7).

*Proof of Theorem 7.* Eq.(8) implies that  $\mathbb{E}_{(\tau_t, a_t) \sim \Gamma^{\pi_b}} [D_{\text{TV}}(M_{\Gamma, \mathbf{b}}(\cdot|\tau_t, a_t), M_{\Gamma}(\cdot|\tau_t, a_t))] \leq \epsilon_1$ . Both this inequality and Eq.(9), through the sub-additivity of TV distance for product distributions, implies

$$D_{\text{TV}}(M_{\Gamma, \mathbf{b}}^{\pi_b}, \Gamma_\star^{\pi_b}) \leq \epsilon_0 H, \quad D_{\text{TV}}(M_{\Gamma, \mathbf{b}}^{\pi_b}, \Gamma^{\pi_b}) \leq \epsilon_1 H.$$

Therefore,  $D_{\text{TV}}(\Gamma^{\pi_b}, \Gamma_\star^{\pi_b}) \leq (\epsilon_0 + \epsilon_1)H$ . Next, we have

$$\mathbb{E}_{(\tau_t, a_t) \sim \Gamma^{\pi_b}} [|Q_{1\text{-RESET}(\Gamma, \mathbf{b})}^{\pi_b}(\tau_t, a_t) - Q_{M_{\Gamma, \mathbf{b}}}^{\pi_b}(\tau_t, a_t)|] \leq \epsilon_1(1 + H)V_{\max},$$

because we can use  $Q_{\Gamma}^{\pi_b}$  as the bridge term, and both  $Q_{1\text{-RESET}(\Gamma, \mathbf{b})}^{\pi_b}$  and  $Q_{M_{\Gamma, \mathbf{b}}}^{\pi_b}$  are close to it thanks to Proposition 4 and Corollary 6. On the other hand, Eq.(9) enables the following through Corollary 6:

$$\mathbb{E}_{(\tau_t, a_t) \sim \Gamma_\star^{\pi_b}} [|Q_{M_{\Gamma, \mathbf{b}}}^{\pi_b}(\tau_t, a_t) - Q_{\Gamma_\star}^{\pi_b}(\tau_t, a_t)|] \leq \epsilon_0 HV_{\max}.$$

Finally, putting everything together:

$$\begin{aligned} & \mathbb{E}_{(\tau_t, a_t) \sim \Gamma_\star^{\pi_b}} [|Q_{1\text{-RESET}(\Gamma, \mathbf{b})}^{\pi_b}(\tau_t, a_t) - Q_{\Gamma_\star}^{\pi_b}(\tau_t, a_t)|] \\ & \leq \mathbb{E}_{(\tau_t, a_t) \sim \Gamma_\star^{\pi_b}} [|Q_{1\text{-RESET}(\Gamma, \mathbf{b})}^{\pi_b}(\tau_t, a_t) - Q_{M_{\Gamma, \mathbf{b}}}^{\pi_b}(\tau_t, a_t)|] + \epsilon_0 HV_{\max} \\ & \leq \mathbb{E}_{(\tau_t, a_t) \sim \Gamma^{\pi_b}} [|Q_{1\text{-RESET}(\Gamma, \mathbf{b})}^{\pi_b}(\tau_t, a_t) - Q_{M_{\Gamma, \mathbf{b}}}^{\pi_b}(\tau_t, a_t)|] + (2\epsilon_0 + \epsilon_1)HV_{\max} \\ & \leq (2\epsilon_0 + 3\epsilon_1)HV_{\max}, \end{aligned}$$

where the third line changes the distribution from  $\Gamma_\star^{\pi_b}$  to  $\Gamma^{\pi_b}$  by paying the for the TV-distance multiplied by the boundedness of the function.  $\square$

9. There is a slight caveat in Stage 2's guarantee due to the violation of realizability, i.e., after stage 1, the true belief state of  $\Gamma_{k^*} = \Gamma_\star$  might have been eliminated. See Appendix D for further discussions.



## 6. Further Discussions

### 6.1 Generalization to Other Sampling Distributions

The roll-out guarantees in Theorem 1 and Corollary 2 all consider Q-function errors under the distribution  $(\tau_t, a_t) \sim \Gamma^{\pi_b}$ , under which we train the classifiers to select the belief-state approximation, and Corollary 2 further restricts the roll-out policy to be  $\pi = \pi_b$ . Naturally, one would wonder what happens when the error is measured under a different sampling distribution (e.g., the occupancy induced from a different roll-in policy  $\pi'$ ), and when REPEATED-RESET is given a general roll-out policy  $\pi \neq \pi_b$ . These questions are well-understood in the MDP literature (especially offline RL theory) that we can pay some form of *coverage coefficient* to translate the error from one distribution to another:

**Proposition 8.** *Consider an MDP with transition  $P$ .*

1. (SINGLE-RESET, extension of Proposition 4) *Given any  $Q : \mathcal{S}_t \times \mathcal{A} \rightarrow \mathbb{R}$  where  $\mathbb{E}_{(s_t, a_t) \sim P^{\pi_b}} [|Q(s_t, a_t) - Q_P^\pi(s_t, a_t)|] \leq \epsilon$  for some fixed  $t$ , for any roll-in policy  $\pi'$ ,*

$$\mathbb{E}_{(s_t, a_t) \sim P^{\pi'}} [|Q(s_t, a_t) - Q_P^\pi(s_t, a_t)|] \leq \epsilon \cdot \left\| \rho_t^{\pi'} / \rho_t^{\pi_b} \right\|_\infty, \quad (10)$$

where  $\rho_t^{(\cdot)}$  is the marginal distribution (a.k.a. occupancy) of  $(s_t, a_t)$  induced by a policy in  $P$ .

2. (REPEATED-RESET, extension of Proposition 2) *Consider another MDP with transition  $P'$  where Eq.(6) holds for all  $t$ . Let  $(\pi')^t \circ (\pi)^{t'-t}$  be the policy that follows  $\pi'$  for the first  $t$  steps and  $\pi$  for the next  $t' - t$  steps, then*

$$\mathbb{E}_{(s_t, a_t) \sim P^{\pi'}} [|Q_{P'}^\pi(s_t, a_t) - Q_P^\pi(s_t, a_t)|] \leq \epsilon \cdot \sum_{t'=t}^{H-1} \left\| \rho_t^{(\pi')^t \circ (\pi)^{t'-t}} / \rho_{t'}^{\pi_b} \right\|_\infty.$$

These results can be directly applied to POMDPs by mapping state  $s_t$  in the proposition to the observable history  $\tau_t$ ,  $P$  to the observable dynamics of  $\Gamma$  (i.e.,  $M_{\Gamma, \mathbf{b}^*}$ ),  $Q$  to  $Q_{1-\text{RESET}(\Gamma, \mathbf{b})}^\pi$ , and  $P'$  to  $M_{\Gamma, \mathbf{b}}$ . However, while the coverage coefficients that appear in the proposition are often acceptable in MDPs, their behaviors are not as benign in POMDPs: for example, Eq.(10) becomes

$$\mathbb{E}_{(\tau_t, a_t) \sim \Gamma^{\pi'}} [|Q_{1-\text{RESET}(\Gamma, \mathbf{b})}^\pi(\tau_t, a_t) - Q_\Gamma^\pi(\tau_t, a_t)|] \leq \epsilon \cdot \left( \max_{\tau_t, a_t} \frac{\mathbb{P}_\Gamma^{\pi'}[\tau_t, a_t]}{\mathbb{P}_\Gamma^{\pi_b}[\tau_t, a_t]} \right), \quad (11)$$

where  $\mathbb{P}_\Gamma^{\pi'}[\tau_t, a_t]$  is the probability assigned to the partial trajectory  $(\tau_t, a_t)$  in  $\Gamma$  under  $\pi'$  as the sampling policy, and

$$\frac{\mathbb{P}_\Gamma^{\pi'}[\tau_t, a_t]}{\mathbb{P}_\Gamma^{\pi_b}[\tau_t, a_t]} = \prod_{t'=0}^t \frac{\pi'(a_{t'} | \tau_{t'})}{\pi_b(a_{t'} | \tau_{t'})}$$

is the infamous cumulative product of importance weights found in importance sampling (Precup et al., 2000). This is actually a general problem whenever we apply MDP results to POMDPs via a reduction to history-based MDPs, and circumventing it often requires

algorithms and coverage concepts specifically designed for POMDPs (Zhang and Jiang, 2024). It remains an interesting question whether those ideas (such as the notion of belief & outcome coverage proposed by Zhang and Jiang (2024)) are useful for the belief-state selection problem considered in this paper.

**Generalization via Sufficient Statistics** A mitigation to the above problem is to make and leverage structural assumptions on  $\mathbf{b}$ . In particular, we may assume that  $\mathbf{b}(\cdot|\tau_t)$  is generated via a two-stage procedure:

$$s_t \sim \mathbf{b}(\cdot|\tau_t) \Leftrightarrow z_t = \phi_{\mathbf{b}}(\tau_t), s_t = \mathbf{b}(\cdot|z_t).$$

That is, we first compute the *sufficient statistic* of  $\tau_t$  via a function  $\phi_{\mathbf{b}}$ , and then sample  $s_t$  conditioned on  $z_t$ . (With a slight abuse of notation we reuse  $\mathbf{b}$  for the conditional distribution in the second stage.) As a starter, if  $z_t$  is a discrete variable and the correct  $\phi_{\mathbf{b}^*}$  is known, i.e.,  $\phi_{\mathbf{b}} = \phi_{\mathbf{b}^*} \forall \mathbf{b} \in \mathcal{B}$ , we can improve the guarantee of LATENT STATE-BASED SELECTION in Eq.(11) to the following (see proof in Appendix C):

$$\mathbb{E}_{(\tau_t, a_t) \sim \Gamma^{\pi'}} [|Q_{1-\text{RESET}(\Gamma, \mathbf{b})}^{\pi}(\tau_t, a_t) - Q_{\Gamma}^{\pi}(\tau_t, a_t)|] \leq \epsilon \cdot \left( \max_{z_t} \frac{\mathbb{P}_{\Gamma}^{\pi'}[z_t]}{\mathbb{P}_{\Gamma}^{\pi_b}[z_t]} \right).$$

Therefore, if  $z_t$  takes on a small number of values, there is hope that  $\pi_b$  may induce an exploratory distribution over  $z_t$  and covers the distribution under  $\pi'$ . The result can also be easily extended to the case of unknown  $\phi_{\mathbf{b}^*}$  (i.e.,  $\phi_{\mathbf{b}}$  can be different for each  $\mathbf{b}$ ), as we can simply replace  $z_t$  in the above bound with the pair  $(\phi_{\mathbf{b}}(\tau_t), \phi_{\mathbf{b}^*}(\tau_t))$ . In this case, the bound requires  $\pi_b$  to induce an exploration *joint* distribution over the pair of statistics. For continuous-valued  $\phi_{\mathbf{b}}$ , favorable coverage guarantees might still be obtainable if structural assumptions are imposed on the  $\mathbf{b}(\cdot|\phi_{\mathbf{b}})$  process.

**Task-specific Approaches** Another route to circumvent the issue related to coverage is to take approaches specific to the task at hand. As an example, for the most basic task of policy evaluation (i.e., estimating  $Q_{\Gamma}^{\pi}$  for a given  $\pi$ ), there are simple regression based methods<sup>10</sup> and selection algorithms based on estimating some variant of the Bellman error. For the latter, the coverage guarantee often does not depend on the coverage in the original MDP, but in an MDP compressed through a low-dimensional representation related to the candidate Q-functions (Xie and Jiang, 2021; Zhang and Jiang, 2021; Liu et al., 2025). Such a deviation from the original dynamics may be a desirable property for POMDPs when the coverage in the original dynamics is not well-behaved.

## 6.2 The Choice of $\pi_b$ : How to Collect Data

For most part of the paper, we assume the  $\pi_b$ , which is used to collect the data needed for the selection of the conditional distributions (Section 3), is given. In practice, the choice of  $\pi_b$  is an important hyperparameter with nuanced effects, which we already had a glimpse

10. That is, we can generate trajectories in  $\Gamma$  with  $\pi'$  roll-in at time step  $t$  and  $\pi$  roll-out, and split each trajectory into a regression data point  $(\tau_t, a_t) \mapsto \sum_{t' \geq t} r_{t'}$ .  $Q_{1-\text{RESET}(\Gamma, \mathbf{b})}^{\pi}$  and  $Q_{M_{\Gamma, \mathbf{b}}}^{\pi}$  are treated as candidate regressors, and the true  $Q_{\Gamma}^{\pi}$  has the least mean squared error.

in Section 5.3: while we choose to select belief-state approximations in each simulator  $\Gamma$  by using the same policy  $\pi_b$  as the one used to sample the real-system data, the analysis needs to handle the mismatch between the roll-in distributions of  $\Gamma^{\pi_b}$  and  $\Gamma_{\star}^{\pi_b}$ , which shows up in the final error bound. While this mismatch is shown to be controlled by  $(\epsilon_0 + \epsilon_1)$ , there is the possibility of using a different roll-in policy  $\pi_{b'}$  in simulators such that  $\Gamma^{\pi_{b'}}$  may be a better approximation of  $\Gamma_{\star}^{\pi_b}$  than  $\Gamma^{\pi_b}$  is. The issue is further complicated when there is misspecification in  $\{\Gamma_k\}$  and  $\mathcal{B}$ , which we leave for future investigation.

Another important motivating scenario for selecting  $\mathbf{b}$  is to use it for learning a good policy in the simulator. In this case, we want  $\mathbf{b}$  to be accurate not just under some fixed distributions, but throughout the learning process when we explore using different policies. A naïve approach is to separately optimize one policy for each candidate  $\mathbf{b} \in \mathcal{B}$ , and rolling out these policies in the simulator to find the best performing one. However, given the computational intensity of policy optimization, an interesting question is whether we can adjust the choice of  $\mathbf{b}$  as policy optimization unfolds and avoid performing  $m = |\mathcal{B}|$  separate policy optimization processes.

## Acknowledgments

The author thanks Akshay Krishnamurthy for valuable discussions on early ideas of the project, Sivaraman Balakrishnan for pointers to relevant work on conditional density estimation, and Preetum Nakkiran and Sam Power for helpful discussions and suggestions related to Bayesian inference.

## References

- Blair Bilodeau, Dylan J Foster, and Daniel M Roy. Minimax rates for conditional density estimation via empirical entropy. *EThe Annals of Statistics*, 51(2):762–790, 2023.
- Cameron B Browne, Edward Powley, Daniel Whitehouse, Simon M Lucas, Peter I Cowling, Philipp Rohlfshagen, Stephen Tavener, Diego Perez, Spyridon Samothrakis, and Simon Colton. A survey of Monte Carlo tree search methods. *IEEE Transactions on Computational Intelligence and AI in Games*, 4(1):1–43, 2012.
- Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *EProceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.
- Luc Devroye and Gábor Lugosi. *ECombinatorial methods in density estimation*. Springer Science & Business Media, 2001.
- Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. Go-explore: a new approach for hard-exploration problems. *EarXiv preprint arXiv:1901.10995*, 2019.
- Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *EThe journal of machine learning research*, 13(1):723–773, 2012.

- Nan Jiang. On value functions and the agent-environment boundary. EarXiv preprint arXiv:1905.13341, 2019.
- Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *EMachine Learning: ECML 2006*, pages 282–293. Springer Berlin Heidelberg, 2006.
- Michael Li, Matey Neykov, and Sivaraman Balakrishnan. Minimax optimal conditional density estimation under total variation smoothness. *Electronic Journal of Statistics*, 16(2):3937–3972, 2022.
- Michael L Littman and Richard S Sutton. Predictive representations of state. In *Advances in neural information processing systems*, pages 1555–1561, 2002.
- Pai Liu, Lingfeng Zhao, Shivangi Agarwal, Jinghan Liu, Audrey Huang, Philip Amortila, and Nan Jiang. Model selection for off-policy evaluation: New algorithms and experimental protocol. EarXiv preprint arXiv:2502.08021, 2025.
- Qinghua Liu, Praneeth Netrapalli, Csaba Szepesvari, and Chi Jin. Optimistic mle—a generic model-based algorithm for partially observable sequential decision making. EarXiv preprint arXiv:2209.14997, 2022.
- Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International conference on artificial intelligence and statistics*, pages 343–351. PMLR, 2021.
- Doina Precup, Richard S Sutton, and Satinder P Singh. Eligibility traces for off-policy policy evaluation. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 759–766, 2000.
- Touqir Sajed, Wesley Chung, and Martha White. High-confidence error estimates for learned value functions. EarXiv preprint arXiv:1808.09127, 2018.
- David Silver and Joel Veness. Monte-Carlo planning in large POMDPs. *Advances in Neural Information Processing Systems*, 23:2164–2172, 2010.
- Satinder Singh, Michael R James, and Matthew R Rudary. Predictive state representations: A new theory for modeling dynamical systems. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 512–519. AUAI Press, 2004.
- Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems. *Journal of Machine Learning Research*, 23(12):1–83, 2022.
- Wen Sun, Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, and John Langford. Model-based RL in Contextual Decision Processes: PAC bounds and Exponential Improvements over Model-free Approaches. In *Conference on Learning Theory*, 2019.
- Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

- Luca Weihs, Unnat Jain, Iou-Jen Liu, Jordi Salvador, Svetlana Lazebnik, Aniruddha Kembhavi, and Alex Schwing. Bridging the imitation gap by adaptive insubordination. *EAdvances in Neural Information Processing Systems*, 34:19134–19146, 2021.
- Tengyang Xie and Nan Jiang. Batch value-function approximation with only realizability. In *International Conference on Machine Learning*, pages 11404–11413. PMLR, 2021.
- Siyuan Zhang and Nan Jiang. Towards hyperparameter-free policy selection for offline reinforcement learning. *EAdvances in Neural Information Processing Systems*, 34:12864–12875, 2021.
- Yuheng Zhang and Nan Jiang. On the curses of future and history in future-dependent value functions for off-policy evaluation. *EAdvances in Neural Information Processing Systems*, 37:124756–124790, 2024.
- Yuheng Zhang and Nan Jiang. Statistical tractability of off-policy evaluation of history-dependent policies in pomdps. *EarXiv preprint arXiv:2503.01134*, 2025.

## Appendix A. Proof of Section 3

*Proof of Theorem 1.* In the proof it suffices to only consider the comparison when  $i^* \in \{i, k\}$ . Under standard concentration argument, the excess risk of ERM on 0/1 classification can be bounded. That is, with the  $N$  given in the theorem statement, under the high-probability event we have the following (the union bound is reflected by the logarithmic dependence on  $n, m, |\mathcal{F}|$  in the sample complexity): let  $\epsilon' = \alpha\epsilon$ ; for any  $i \neq i^*$ ,

$$\text{acc}_{X_j}^{i, i^*}(\hat{f}_j^{i, i^*}) \geq \text{acc}_{X_j}^{i, i^*}(f_{X_j}^{i, i^*}) - \epsilon'/24. \quad (12)$$

We now link  $\text{acc}_X^{i, i^*}(f)$  to the discrimination power of  $f$  w.r.t.  $P_i$  and  $P^*$ : given that  $f$  has binary output, we have  $\Pr[f = 1] = \mathbb{E}[f]$ , and

$$\begin{aligned} \text{acc}_X^{i, i^*}(f) &= 1/2 \cdot \left( \Pr_{Y \sim P^*(\cdot|X)}[f(Y) = 1] + 1 - \Pr_{Y \sim P_i(\cdot|X)}[f(Y) = 1] \right) \\ &= 1/2 + 1/2 \cdot (\mathbb{E}_{Y \sim P^*(\cdot|X)}[f(Y)] - \mathbb{E}_{Y \sim P_i(\cdot|X)}[f(Y)]). \end{aligned} \quad (13)$$

Replacing  $\text{acc}_X^{i, i^*}(f)$  in Eq.(12) with the above expression, we have  $\forall j$ ,

$$\begin{aligned} &\mathbb{E}_{Y \sim P^*(\cdot|X_j)}[\hat{f}_j^{i, i^*}(Y)] - \mathbb{E}_{Y \sim P_i(\cdot|X_j)}[\hat{f}_j^{i, i^*}(Y)] \\ &\geq \mathbb{E}_{Y \sim P^*(\cdot|X_j)}[f_{X_j}^{i, i^*}(Y)] - \mathbb{E}_{Y \sim P_i(\cdot|X_j)}[f_{X_j}^{i, i^*}(Y)] - \epsilon'/12. \end{aligned} \quad (14)$$

Now we consider the concentration of empirical averages to their (conditional) expectations in the final scoring rule in Eq.(1) when  $i^* \in \{i, k\}$ : Conditioned on  $\{X_j\}_{j=1}^n$  and all the synthetic data drawn to train the classifiers (which are independent to the randomness of drawing  $Y_j$  given  $X_j$ ),  $\frac{1}{n} \sum_j \hat{f}_j^{i, k}(Y_j)$  is the average of independent (but generally not identically distributed) random variables, each of which has conditional mean  $\mathbb{E}_{Y \sim P^*(\cdot|X_j)}[\hat{f}_j^{i, k}(Y)]$ . Therefore, by Hoeffding's inequality, the  $n$  in the statement enables that

$$\left| \frac{1}{n} \sum_j \hat{f}_j^{i, k}(Y_j) - \frac{1}{n} \sum_j \mathbb{E}_{Y \sim P^*(\cdot|X_j)}[\hat{f}_j^{i, k}(Y)] \right| \leq \epsilon'/12. \quad (15)$$

The same argument holds for  $\frac{1}{n} \sum_j \hat{f}_j^{i, k}(Y_j^{(i)})$  since  $Y_j^{(i)}$  is holdout data not used in training:

$$\left| \frac{1}{n} \sum_j \hat{f}_j^{i, k}(Y_j^{(i)}) - \frac{1}{n} \sum_j \mathbb{E}_{Y \sim P_i(\cdot|X_j)}[\hat{f}_j^{i, k}(Y)] \right| \leq \epsilon'/12. \quad (16)$$

We now consider the final score in Eq.(1) when  $i^* \in \{i, k\}$ . First consider  $i = i^*$ : the two averages share the same mean, so the difference is always bounded by  $\epsilon'/6$  given the concentration bounds above regardless of  $k$ . In the second case,  $i \neq i^*$ ,  $k = i^*$ , where the two averages have different means. Using the concentration bounds above, we have

$$\begin{aligned} &\left| \frac{1}{n} \sum_j \hat{f}_j^{i, i^*}(Y_j) - \frac{1}{n} \sum_j \hat{f}_j^{i, i^*}(Y_j^{(i)}) \right| \\ &\geq \frac{1}{n} \sum_j \hat{f}_j^{i, i^*}(Y_j) - \frac{1}{n} \sum_j \hat{f}_j^{i, i^*}(Y_j^{(i)}) \\ &\geq \frac{1}{n} \sum_j \mathbb{E}_{Y \sim P^*(\cdot|X_j)}[\hat{f}_j^{i, i^*}(Y)] - \frac{1}{n} \sum_j \mathbb{E}_{Y \sim P_i(\cdot|X_j)}[\hat{f}_j^{i, i^*}(Y)] - \epsilon'/6. \quad (\text{Eqs. (15) and (16)}) \end{aligned} \quad (17)$$

For the final  $\hat{i}$  being selected, if  $\hat{i} \neq i^*$ , it must be the case that its score is less than that of  $i^*$  which is at most  $\epsilon'/6$ , so Eq.(17) for  $i = \hat{i}$  is at most  $\epsilon'/6$ , and hence

$$\frac{1}{n} \sum_j \mathbb{E}_{Y \sim P^*(\cdot|X_j)}[\hat{f}_j^{\hat{i}, i^*}(Y)] - \frac{1}{n} \sum_j \mathbb{E}_{Y \sim P_i(\cdot|X_j)}[\hat{f}_j^{\hat{i}, i^*}(Y)] \leq \epsilon'/3.$$

Combine this with Eq.(14), we have

$$\frac{1}{n} \sum_j \mathbb{E}_{Y \sim P^*(\cdot|X_j)}[\hat{f}_{X_j}^{\hat{i}, i^*}(Y)] - \frac{1}{n} \sum_j \mathbb{E}_{Y \sim P_i(\cdot|X_j)}[\hat{f}_{X_j}^{\hat{i}, i^*}(Y)] \leq 5\epsilon'/12.$$

The next step is to show concentration for the LHS of the above expression. Note that each term like  $\mathbb{E}_{Y \sim P^*(\cdot|X_j)}[\hat{f}_{X_j}^{\hat{i}, i^*}(Y)]$  is a non-random property of  $X_j$  (we need to union bound over  $i$  so that it applies to  $i = \hat{i}$ ), so their average concentrates to the population expectation with  $\epsilon'/12$  error under the  $n$  in the theorem statement. Putting together,

$$\mathbb{E}_{X \sim P^*} [\mathbb{E}_{Y \sim P^*(\cdot|X)}[\hat{f}_X^{\hat{i}, i^*}(Y)] - \mathbb{E}_{Y \sim P_i(\cdot|X)}[\hat{f}_X^{\hat{i}, i^*}(Y)]] \leq \epsilon'/2.$$

Finally,

$$\begin{aligned} 2\mathbb{E}_{X \sim P^*} [\text{acc}_X^{\hat{i}, i^*}(f_X^{\hat{i}, i^*})] - 1 &= \mathbb{E}_{X \sim P^*} [\mathbb{E}_{Y \sim P^*(\cdot|X)}[\hat{f}_X^{\hat{i}, i^*}(Y)] - \mathbb{E}_{Y \sim P_i(\cdot|X)}[\hat{f}_X^{\hat{i}, i^*}(Y)]] \\ &\quad (\text{Eq.(13)}) \\ &\leq \epsilon'/2. \end{aligned}$$

Given Assumption 2, we have

$$1/2 + \epsilon'/4 \geq \mathbb{E}_{X \sim P^*} [\text{acc}_X^{\hat{i}, i^*}(f_X^{\hat{i}, i^*})] \geq 1/2 + \alpha \cdot \mathcal{E}(\hat{i}, i^*),$$

so  $\mathcal{E}(\hat{i}, i^*) \leq \epsilon'/4\alpha = \epsilon/2$ . The proof is concluded by noticing that

$$\mathcal{E}(\hat{i}, i^*) = 1/2 \cdot \mathbb{E}_{X \sim P^*} [D_{\text{TV}}(P_{\hat{i}}(\cdot|X), P_{i^*}(\cdot|X))].$$

□

## Appendix B. Proof of Section 5

*Proof of Theorem 5.* To avoid dealing with the last time step separately, we take the convention that any notion of value function evaluates to 0 at  $t = H$  since there is no future reward afterwards. Then for any  $t < H$ , when  $\pi = \pi_b$ ,

$$\begin{aligned} &\mathbb{E}_{(s_t, a_t) \sim P^{\pi_b}} [ |Q_P^\pi(s_t, a_t) - Q_{P'}^\pi(s_t, a_t)| ] \\ &= \mathbb{E}_{(s_t, a_t) \sim P^{\pi_b}} [ |E_{s_{t+1} \sim P(\cdot|s_t, a_t)}[R(s_{t+1}) + V_P^\pi(s_{t+1})] - E_{s_{t+1} \sim P'(\cdot|s_t, a_t)}[R(s_{t+1}) + V_{P'}^\pi(s_{t+1})]| ] \\ &\leq \mathbb{E}_{(s_t, a_t) \sim P^{\pi_b}} [ |E_{s_{t+1} \sim P(\cdot|s_t, a_t)}[R(s_{t+1}) + V_P^\pi(s_{t+1})] - E_{s_{t+1} \sim P'(\cdot|s_t, a_t)}[R(s_{t+1}) + V_{P'}^\pi(s_{t+1})]| \\ &\quad + |E_{s_{t+1} \sim P(\cdot|s_t, a_t)}[V_P^\pi(s_{t+1}) - V_{P'}^\pi(s_{t+1})]| ] \\ &\leq \mathbb{E}_{(s_t, a_t) \sim P^{\pi_b}} [D_{\text{TV}}(P(\cdot|s_t, a_t), P'(\cdot|s_t, a_t)) \cdot V_{\max}] \\ &\quad + \mathbb{E}_{(s_t, a_t) \sim P^{\pi_b}} [E_{s_{t+1} \sim P(\cdot|s_t, a_t)}[|V_P^\pi(s_{t+1}) - V_{P'}^\pi(s_{t+1})|]] \\ &\leq \epsilon V_{\max} + \mathbb{E}_{s_{t+1} \sim P^{\pi_b}} [ |Q_P^\pi(s_{t+1}, \pi) - Q_{P'}^\pi(s_{t+1}, \pi)| ] \\ &\leq \epsilon V_{\max} + \mathbb{E}_{(s_{t+1}, a_{t+1}) \sim P^{\pi_b}} [ |Q_P^\pi(s_{t+1}, a_{t+1}) - Q_{P'}^\pi(s_{t+1}, a_{t+1})| ], \end{aligned}$$



where the last step uses the fact that  $\pi = \pi_b$ , and inductively expanding the analysis till the end proves the theorem statement.  $\square$

## Appendix C. Proof of Section 6

**Proposition 9.** *If Eq.(3) holds for some  $t$  and all  $\mathbf{b} \in \mathcal{B}$  share the same sufficient statistics  $\phi$ , then given any roll-in policy  $\pi'$ , we have*

$$\mathbb{E}_{(\tau_t, a_t) \sim \Gamma^{\pi'}} [|Q_{1\text{-RESET}(\Gamma, \mathbf{b})}^{\pi}(\tau_t, a_t) - Q_{\Gamma}^{\pi}(\tau_t, a_t)|] \leq \epsilon \cdot \left( \max_{z_t} \frac{\mathbb{P}_{\Gamma}^{\pi}[z_t]}{\mathbb{P}_{\Gamma}^{\pi_b}[z_t]} \right).$$

*Proof.* Following the proof of Proposition 4, we know that the LHS is controlled by

$$\mathbb{E}_{(\tau_t, a_t) \sim \Gamma^{\pi'}} [D_{\text{TV}}(\mathbf{b}(\cdot|\tau_t), \mathbf{b}^*(\cdot|\tau_t))] = \mathbb{E}_{z_t \sim \Gamma^{\pi'}} [D_{\text{TV}}(\mathbf{b}(\cdot|z_t), \mathbf{b}^*(\cdot|z_t))],$$

where  $z_t = \phi(\tau_t)$ . Similarly, Eq.(3) gives us

$$\epsilon \geq \mathbb{E}_{\tau_t \sim \Gamma^{\pi_b}} [D_{\text{TV}}(\mathbf{b}(\cdot|\tau_t), \mathbf{b}^*(\cdot|\tau_t))] = \mathbb{E}_{z_t \sim \Gamma^{\pi_b}} [D_{\text{TV}}(\mathbf{b}(\cdot|z_t), \mathbf{b}^*(\cdot|z_t))].$$

Performing change of measure w.r.t.  $z_t$  immediately completes the proof.  $\square$

## Appendix D. Discussion of Theorem 7

As mentioned in Footnote 9, to directly apply our analysis for OBSERVATION-BASED SELECTION in the second stage we require realizability, i.e., the  $M_{\Gamma_{\star}} \in \{M_{\Gamma_k, \mathbf{b}_{\Gamma_k}}\}$ , where  $\mathbf{b}_{\Gamma_k}$  is the belief state approximation selected for  $\Gamma_k$ . This does not always hold, because for  $\Gamma_{k^{\star}} = \Gamma_{\star}$ , the selected  $\mathbf{b}_{\Gamma_{k^{\star}}}$  may not be its true belief state, causing the non-realizability.

There are two fixes to this issue, both still leading to the kind of guarantee in Eq.(9): in the first fix, we can extend the analysis in Theorem 1 to handle misspecification. In the second fix, we can change the algorithm as follows:

1. Run Stage 2 with all  $\{(\Gamma, \mathbf{b}) : \Gamma \in \{\Gamma_k\}, \mathbf{b} \in \mathcal{B}\}$ . This way, realizability is satisfied and we obtain the guarantee for OBSERVATION-BASED SELECTION.
2. When running the conditional selection algorithm for both LATENT STATE-BASED SELECTION (Stage 1) and OBSERVATION-BASED SELECTION (Stage 2), we do not take the argmin of the score but the version space, i.e., the set of candidate conditionals that is plausible to be the true conditional. In the proof of Theorem 1, this corresponds to the set of  $P_i$  whose score is no greater than  $\epsilon'/6$  (see the paragraph below Eq.(16)). It is easy to see that all conditional distributions in the version space enjoy the guarantee of Theorem 1.
3. For each  $\Gamma \in \{\Gamma_k\}$ , we pair it with each plausible belief state, and gather such pairs across  $\{\Gamma_k\}$ . Then, we take the intersection between this set and the version space of Stage 2. Any  $(\Gamma, \mathbf{b})$  pair in the intersection must enjoy both the guarantee of OBSERVATION-BASED SELECTION and that of LATENT STATE-BASED SELECTION, thus satisfying the conditions of Theorem 7.