Learning in POMDPs is Sample-Efficient with Hindsight Observability

Jonathan N. Lee [1 2] **Alekh Agarwal** [2] **Christoph Dann** [2] **Tong Zhang** [2 3]

Abstract

POMDPs capture a broad class of decision making problems, but hardness results suggest that learning is intractable even in simple settings due to the inherent partial observability. However, in many realistic problems, more information is either revealed or can be computed during some point of the learning process. Motivated by diverse applications ranging from robotics to data center scheduling, we formulate a Hindsight Observable Markov Decision Process (HOMDP) as a POMDP where the latent states are revealed to the learner in hindsight and only during training. We introduce new algorithms for the tabular and function approximation settings that are provably sample-efficient with hindsight observability, even in POMDPs that would otherwise be statistically intractable. We give a lower bound showing that the tabular algorithm is optimal in its dependence on latent state and observation cardinalities.

1. Introduction

Sequential decision making settings where the learning agent only receives incomplete observations of its environmental state are typical in diverse practical scenarios, such as control of physical systems (Thrun, 2000), dialogue and recommendation systems (Young et al., 2013; Shani et al., 2005), and decision making in educational or clinical settings (Ayer et al., 2012). Typically studied within the framework of a Partially Observable Markov Decision Process (POMDP), classical literature on such problems provides hardness results on sample and computationally efficient learning, even in simple settings with small action, observation and state spaces, in stark contrast to the MDP setting where the state is fully observable. Fueled by this gap, there is a body of literature that characterizes observability conditions when the sequence of observations reveals enough information about the latent states to permit sample-efficient learning. In this paper, we ask if the motivating practical applications sometimes allow a more informative sensing of the underlying state for some part of the learning process. We formulate a novel learning setting called a Hindsight Observable Markov Decision Process (HOMDP), and provide learning algorithms that are significantly more sample-efficient than those for general POMDPs.

For motivation, let us consider robotic control, where we want our robot to sense its state using a relatively cheap camera sensor upon deployment. However, during training, it is common to allow a more expensive sensing of the state, using simulators, higher-fidelity cameras, lidars, or even full-fledged motion capture setups (Pinto et al., 2017; Pan et al., 2017; Chen et al., 2020). In a completely different style of scenarios, Sinclair et al. (2022) discuss problems such as scheduling in a data center, where the unknown lifetime of a job creates partial observability of the state when the job is scheduled. This partial observability is resolved when the job actually concludes. While the two examples are very different, they share a similarity. The learner needs a decision making policy to act based on partial observations alone, due either to resource/sensor constraints upon deployment or to fundamental lack of information at the time of decision. However, the underlying state eventually gets revealed, either intrinsically, or due to extrinsic measurements during the training process. We refer to this eventual observation of the latent environment state as *hindsight observability*, and study learning settings where the learner acts based on partial state observations, but observes the true latent states eventually upon the conclusion of the trajectory.

We start by noting that learning in a HOMDP remains considerably challenging in comparison with MDPs, as the learner's policy needs to depend on observations during deployment (e.g. robotics, scheduling) and sometimes even during training (e.g. scheduling). Hence we cannot use MDP learning techniques directly. At the same time, the HOMDP model eliminates the identifiability or observability conditions that are crucial to success in POMDP learning, since the hindsight observation of the latent state allows us to associate latent states and corresponding observations, albeit with a delay. This makes the HOMDP an intermediate step between the complexity of MDPs and POMDPs, which is practically prevalent as our earlier examples suggest.

[1]Stanford University [2]Google Research [3]HKUST. Correspondence to: Jonathan N. Lee <jnl@stanford.edu>.

Our contributions In addition to formalizing the HOMDP framework, and showing its broad applicability across diverse settings, such as sim-to-real robotics, high-frequency control, meta-learning and scheduling problems in Appendix A, we make the following key contributions:

1. When the latent states and observations are both finite, we provide an algorithm HOP-B, which finds an ϵ-optimal policy using at most $\widetilde{\mathcal{O}}\left(\frac{XYH^5+XAH^4}{\epsilon^2}\right)$ trajectories, where the HOMDP contains X latent states, Y observations, A actions, and the horizon is H. In contrast with standard POMDP learning results, there is no observability-related parameter in our bound.
2. We show an $\Omega(\frac{XY}{\epsilon^2})$ lower bound, meaning that HOP-B scales optimally with latent states and observations.
3. We develop a general algorithm, HOP-V, which allows function approximation for both latent states and observations, and allows representation learning in the latent state space. The sample complexity of HOP-V depends on the statistical complexity of function classes used to learn latent state transitions and emissions, along with a rank parameter of the latent state transitions. Again, there are no observability conditions in contrast with standard POMDP results.

2. Related Work

There has been significant progress in understanding the sample efficiency of reinforcement learning in the fully observable setting of MDPs. For tabular MDPs (finite states and actions), upper and lower bounds for sample complexity and regret are well known (Auer et al., 2008; Dann & Brunskill, 2015; Osband & Van Roy, 2016; Azar et al., 2017; Dann et al., 2019; Zanette & Brunskill, 2019). Similar results have been established for MDPs that satisfy certain structural conditions, enabling function approximation (Jiang et al., 2017; Sun et al., 2019; Jin et al., 2020b; Agarwal et al., 2020; Du et al., 2021; Jin et al., 2021a; Foster et al., 2021; Agarwal & Zhang, 2022).

Relative to MDPs, the sample complexity of reinforcement learning in POMDPs is less understood. Classical hardness results suggest learning in POMDPs can be both computationally and statistical intractable even for simple settings (Krishnamurthy et al., 2016). This hardness has spurred researchers to identify conditions under which sample efficient learning is still possible in POMDPs. Block MDPs (Krishnamurthy et al., 2016; Du et al., 2019) and decodable MDPs (Efroni et al., 2022) are special classes of POMDPs in which the current observation (or last few observations) can exactly decode the current latent state. Several works study more general observability conditions beyond decodability (Azizzadenesheli et al., 2016; Guo et al., 2016; Jin et al., 2020a; Golowich et al., 2022; Liu et al., 2022a;b; Uehara et al., 2022; Chen et al., 2022). Sample complexity bounds under these conditions often depend crucially on parameters that quantify the degree of observability. Liu et al. (2022b); Zhan et al. (2022); Zhong et al. (2022) provide similar conditions for general predictive state representations (PSRs). Although aimed at the same objective of learning policies for partially observable settings, our work uses hindsight observability to circumvent any additional parameters or assumptions on the emission function.

Empirically, a number of works successfully leverage latent state information during training to improve sample efficiency. Pinto et al. (2017); Baisero & Amato (2021) study *asymmetric* actor-critic algorithms where the critic uses the latent state while the actor uses observations, allowing the learned policy to later interact with only observations. Pan et al. (2017); Chen et al. (2020); Warrington et al. (2021) use distillation-based approaches where they train an expert policy on latent states and then later imitate it with an observation-based policy. Similar settings also appear as privileged information (Kamienny et al., 2020) or resource-constrained RL (Regatti et al., 2021). However, these prior works do not address sample complexity and exploration.

Motivated by resource allocation, Sinclair et al. (2022) study a similar hindsight problem, where the unobserved part of the latent state is not affected by the learner's actions, and dynamics are fully known in hindsight. Hence, they study a planning problem in hindsight with no need for exploration, unlike the general HOMDP setting considered here.

Kwon et al. (2021); Zhou et al. (2022) study a latent MDP model, where the latent state contains an additional identifier of the active MDP for each episode, and the identifier is revealed in hindsight during training. Our setting is significantly more general, but shares similar motivation. We compare our bounds in Section 4.

3. Hindsight Observable Markov Decision Process

The underlying model of the HOMDP setting is the same as a POMDP; the difference lies in what information is revealed to the learner and when. We first review the relevant quantities of a POMDP and subsequently introduce the hindsight observability and learning protocol in a HOMDP.

3.1. Preliminaries

For $n \in \mathbb{N}$, we use $[n]$ to denote the set $\{1, 2, \ldots, n\}$. For a set S, $\Delta(S)$ denotes the set of (appropriately defined) densities over S. For $h \in \mathbb{N}$, we use $a_{1:h}$ to denote $(a_1, \ldots, a_h)$.

We consider an episodic partially observable Markov decision process (POMDP) $\mathcal{M}$ with episode length H, latent state space $\mathcal{X}$, observation space $\mathcal{Y}$, and action space

$\mathcal{A}$. When these are finite, we denote their respective cardinalities as $X := |\mathcal{X}|$, $Y := |\mathcal{Y}|$, and $A := |\mathcal{A}|$. An initial latent state x_1 is sampled from a fixed and known initial state distribution $\rho \in \Delta(\mathcal{X})$. The process evolves according to transition function $T_\star : \mathcal{X} \times \mathcal{A} \to \Delta(\mathcal{X})$, acting on the latent states. When the learner visits a latent state x, the environment generates an observation $y \in \mathcal{Y}$ according to the conditional emission function $O_\star : \mathcal{X} \to \Delta(\mathcal{Y})$. In particular, in each episode, a (latent) trajectory $\bar{\tau} = (x_1, y_1, a_1, \ldots, x_H, y_H, a_H, x_{H+1})$ is generated where $x_1 \sim \rho(\cdot)$, $x_{h+1} \sim T_\star(\cdot|x_h, a_h)$, $y_h \sim O_\star(\cdot|x_h)$, and the learner selects $a_{1:H}$. We include x_{H+1} (from taking a_H in x_H) as a latent variable for convenience. When referring to an (observed) trajectory of just observations and actions, we use $\tau = (y_1, a_1, \ldots, y_H, a_H)$. We assume there is a known deterministic reward function $r : \mathcal{X} \times \mathcal{A} \to [0, 1]$. Our results can be generalized readily to stochastic, observation-dependent rewards.

As is standard in POMDPs, we consider the setting where the learner interacts with the environment by specifying a *history-dependent* policy $\pi : (\mathcal{Y}\times\mathcal{A})^* \times \mathcal{Y} \to \Delta(\mathcal{A})$ which takes as input a (variable) h-length history of observations $y_{1:h}$ and $(h-1)$-length history of actions $a_{1:h-1}$ and outputs a distribution over actions. That is, the learner's policy does not get to observe any of the latent states $x_{1:H+1}$ during execution of π. For conciseness, we denote the partial histories as $\tau_h := (y_{1:h}, a_{1:h-1})$ and $\bar{\tau}_h := (y_{1:h}, x_{1:h}, a_{1:h-1})$, which includes the observation y_h and latent state x_h (if applicable) at step h.

For a policy π, we denote the expected cumulative reward over an episode by

$$v(\pi) = \mathbb{E}_\pi \left[\sum_{h\in[H]} r(x_h, a_h) \right], \tag{1}$$

where $\mathbb{E}_\pi$ is the expectation taken over trajectories in the POMDP under policy π.

3.2. Hindsight observability

Now we formally introduce the HOMDP setting and describe the interaction protocol, i.e., how the learner interacts with the environment and receives information. We also illustrate this description in the accompanying Figure 1. Along the way, we highlight differences with the standard POMDP and MDP settings. There are two phases in the HOMDP: train time and test time.

During train time, the learner interacts with the environment over $K \in \mathbb{N}$ rounds (episodes). At any given round $k \in [K]$, the learner produces a history-dependent policy $\hat{\pi}_k$ which is deployed in the partially observable environment as if the learner is interacting with a standard POMDP. During execution of the episode k at time $h \in [H]$, the environment reveals only the current observation y_h to the learner. The policy can thus base its decision a_h on only the partial history $\tau_h = (y_{1:h}, a_{1:h-1})$ of interactions in that episode. Once the kth episode is completed, the latent states $x_{1:H+1}$ are revealed to the learner in hindsight, hence the terminology *hindsight observability*. The learner can then generate a new policy $\hat{\pi}_{k+1}$ using information from $(x_{1:H+1}, y_{1:H}, a_{1:H})$ as well as that of all previous episodes. This is the key difference between HOMDPs and standard POMDPs where the latent states are never revealed and the learner generates the policy from only previous observations, actions, and rewards alone. In MDPs, on the other hand, the latent state is observed instantaneously and the policy can directly map a latent state to an action.

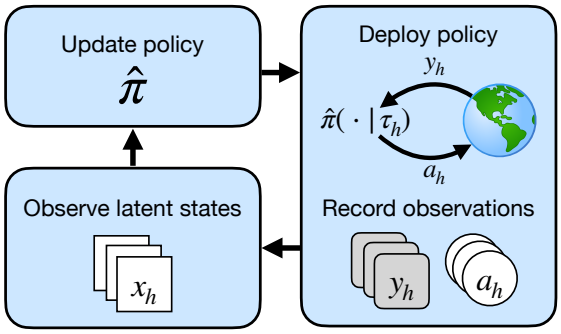

Figure 1: A HOMDP model at train time. The learned history-dependent policy $\hat{\pi}$ is deployed and takes actions $a_{1:H}$ using only observations $y_{1:H}$. After deployment, the environment reveals the latent states $x_{1:H+1}$. The policy updates with both the latent states $x_{1:H+1}$ and observations $(y_{1:H}, a_{1:H})$. At test time, only a history-dependent policy is deployed.

The train time phase may be followed by a test time phase where a single history-dependent policy $\hat{\pi}$ is deployed but the learner does not observe latent states or update the policy after committing to $\hat{\pi}$. To determine $\hat{\pi}$, the learner can use all of the information collected over the K episodes at train time, including the latent states observed in hindsight. The quantity $v(\hat{\pi})$ evaluates the quality of $\hat{\pi}$. We let $\pi^\star$ denote the optimal observation-based policy maximizing $v(\pi^\star)$ and measure the sub-optimality of the learner's policy $\hat{\pi}$ by the difference $v(\pi^\star) - v(\hat{\pi})$. Again, in contrast with an MDP, a HOMDP never reveals the latent state at test time.

We are primarily concerned with PAC sample complexity bounds controlling the suboptimality of $\hat{\pi}$, but some algorithms also address the regret problem at train time where the regret for all K rounds is measured as

$$\mathrm{Reg}(K) = \sum_{k\in[K]} v(\pi^\star) - v(\hat{\pi}_k).$$

Example 1 (Sim-to-real robotics)**.** In *sim-to-real robotics* (Pinto et al., 2017), one trains an image-based policy in a simulator with access to the underlying states. The goal is to deploy the image-based policy in the real world. $\mathcal{X}$ is the set

of robot and object positions and poses which are observable in the simulator during training. $\mathcal{Y}$ is the set of image observations from a camera, which is the only modality available at test time. $\mathcal{X}$ and $\mathcal{Y}$ are both continuous and high dimensional. $\mathcal{A}$ is the set of control inputs the robot can take such as joint torques. Note that the latent states are available without delay at train time in this example. However, we later show that this variant of the problem interestingly does not yield significant statistical advantage in theory because the desired policy at test time is still history-dependent (see the discussion following Theorem 5.1). Empirically, history-dependent policies are still preferred to state-based policies even during train time despite access to the state to facilitate better sim-to-real transfer.

Example 2 (Data center scheduling)**.** In *data center scheduling* (Sinclair et al., 2022), described in the introduction, $\mathcal{Y}$ is the observable state of the submitted, processing, and completed jobs as well as their allocations to servers, which is available at the time of decision-making. $\mathcal{X}$ is a concatenation of $\mathcal{Y}$ with lifetime lengths of the submitted/processing jobs and the arrival times of future jobs. This information is available, but only in hindsight. Depending on the setup, $\mathcal{X}$ and $\mathcal{Y}$ can be relatively succinct here. $\mathcal{A}$ is the set of allocation actions for currently submitted jobs.

3.3. Comparison with hardness of learning in POMDPs

Both POMDPs and HOMDPs share the use of history-dependent policies during execution at test time. However, a POMDP never reveals the association between observations and latent states, leading to a lack of identifiability and exponential in H lower bounds even for simple ones (Krishnamurthy et al., 2016). As discussed previously, numerous recent papers (Liu et al., 2022a; Jin et al., 2020a; Cai et al., 2022; Liu et al., 2022b; Zhan et al., 2022) investigate observability conditions under which sample-efficient learning is possible by ensuring $O_\star$ reveals enough about the distribution of possible latent states. This yields sample complexity bounds that incur an unavoidable dependence on the minimum singular value of $O_\star$ (Liu et al., 2022a), or related parameters.

However, settings where such observability conditions are satisfied may still preclude many practically interesting partially observable problems. Our objective in this paper is to understand to what extent the addition of hindsight observability in a HOMDP can make learning in partially observable settings sample-efficient without relying on observability parameters.

4. Learning in Finite HOMDPs

We now turn to the design of efficient algorithms for learning in the HOMDP model. We begin with the setting where the latent state space $\mathcal{X}$ and observation space $\mathcal{Y}$ have finite cardinalities $X := |\mathcal{X}|$ and $Y := |\mathcal{Y}|$. We introduce a new algorithm, HOP-B, which naturally extends minimax optimal results (in X and A) for tabular MDPs to the HOMDP model. Our proposed algorithm is model-based, leveraging the intuition that, provided with the latent states $x_{1:H+1}$, one should be able to estimate the transition and emission functions, $T_\star$ and $O_\star$. We start with the algorithm, before presenting the sample complexity guarantee.

4.1. The HOP-B algorithm

Our algorithm, Hindsight OPtimism with Bonus (HOP-B) estimates the transition and emission models, and subsequently finds an optimal policy in this learned model using a reward bonus to encourage exploration. The design of bonus is a key novelty in HOP-B, relative to its MDP counterparts, as we will discuss shortly.

Before describing the algorithm, we define a planning oracle which is used in the algorithm to compute the exploration policies. Note that planning in a HOMDP is identical to a POMDP, as we seek an optimal history-dependent policy.

Definition 4.1 (Optimal planner)**.** The POMDP planner POP takes as input a transition function T, an emission function O, and a reward function r and returns a policy $\pi = \mathsf{POP}(T, O, r)$ such that $v(\pi) = \max_{\pi'} v_{\mathcal{M}(T,O,r)}(\pi')$, where $\mathcal{M}(T, O, r)$ denotes the POMDP model with latent transitions T, emissions O, and reward function r.

While it is known that planning in POMDPs is PSPACE-hard in general (Papadimitriou & Tsitsiklis, 1987), there are many special classes of POMDPs for which planning is computationally efficient. Alternatively, it is possible in practice to use one of many existing approximate POMDP planners; however, this will likely weaken the subsequent theoretical guarantees of this section up to some approximation error. Regardless, this is much milder computational assumption than what is sometimes made in comparable POMDP literature (Jin et al., 2020a).

HOP-B operates over K rounds, starting with arbitrary guesses $\hat{T}_1$ and $\hat{O}_1$ of the model. At round k, it computes reward bonuses based on the uncertainty in the estimates $\hat{T}_k$ and $\hat{O}_k$, which is quantified by the number of visits to each latent state x and latent-state action pair (x, a) from the dataset. We define these bonuses in $\epsilon_k(x)$ and $\epsilon_k(x, a)$ in lines 7 and 8 with parameters given in line 4. Note that the $\epsilon_k(x)$ bonus is in addition to the typical bonus in MDPs. Informally $\epsilon_k(x, a)$ captures our uncertainty in the estimation of $T_\star$, while $\epsilon_k(x)$ measures it for $O_\star$. For instance, even if we know $T_\star$, we need to visit each latent state to estimate its emission process for the subsequent planning, capturing the need for the additional $\epsilon_k(x)$ bonus.

We construct a reward function $\hat{r}_k$ by adding the bonuses to r. We then invoke the planner POP using $\hat{T}_k$, $\hat{O}_k$, and

Algorithm 1 Hindsight OPtimism with Bonus (HOP-B)

Input: POMDP planner POP.
Initialize emission and transition models $\hat{O}_1, \hat{T}_1$.
Initialize $n_1(x) = n_1(x,a) = 0$ for all $x \in \mathcal{X}, a \in \mathcal{A}$.
Set bonus parameters $\beta_1 = 4H^3 \log(YXAHK/\delta), \beta_2 = 8Y \log(YXKH/\delta)$.
for $k = 1, \dots, K$ **do**
  `// Set reward bonuses`
  $\epsilon_k(x,a) = \min\left\{\sqrt{\frac{\beta_1}{n_k(x,a)}}, 2H\right\}$
  $\epsilon_k(x) = \min\left\{\sqrt{\frac{\beta_2}{n_k(x)}}, 2\right\}$
  $\hat{r}_k(x,a) = r(x,a) + H\epsilon_k(x) + \epsilon_k(x,a)$
  `// Plan, deploy hist.-dependent policy`
  $\hat{\pi}_k = \mathsf{POP}(\hat{T}_k, \hat{O}_k, \hat{r}_k)$
  Run $\hat{\pi}_k$ and observe trajectory $\tau^k = (y^k_{1:H}, a^k_{1:H})$.
  `// Hindsight observation`
  Observe latent states $x^k_{1:H+1} = (x^k_1, \dots, x^k_{H+1})$.
  `// Update models`
  $n_{k+1}(x) = \sum_{\ell \in [k], h \in [H]} \mathbf{1}\{x^\ell_h = x\}$.
  $n_{k+1}(x,a) = \sum_{\ell \in [k], h \in [H]} \mathbf{1}\{x^\ell_h = x \wedge a^\ell_h = a\}$.
  Update $\hat{T}_{k+1}$ via (2)
  Update $\hat{O}_{k+1}$ via (3)
end for

the reward function $\hat{r}_k$ to generate an optimistic history-dependent policy $\hat{\pi}_k$. As we show in the proof, the estimated value of $\hat{\pi}_k$ under the current model over-estimates the true value of $\pi^\star$ with high probability. We then deploy the optimistic policy $\hat{\pi}_k$ in the environment to generate a trajectory of observations $y_{1:H}$ and actions $a_{1:H}$. We further observe the latent states $x_{1:H+1}$ in hindsight. Finally, using the new information from the trajectory and in hindsight, we update the models with empirical estimates using all the past data:

$$\hat{T}_{k+1}(x'|x,a) = \sum_{\ell \in [k], h \in [H]} \frac{\mathbf{1}\left\{x^\ell_h = x, y^\ell_h = y, x^\ell_{h+1} = x'\right\}}{n_{k+1}(x,a)} \quad (2)$$

$$\hat{O}_{k+1}(y|x) = \sum_{\ell \in [k], h \in [H]} \frac{\mathbf{1}\left\{x^\ell_h = x, y^\ell_h = y\right\}}{n_{k+1}(x)} \quad (3)$$

for all x, x', a, y where $n_{k+1}(x,a)$ and $n_{k+1}(x)$ are defined in Algorithm 1. Note that both the calculation of the uncertainty bonuses and the model updates are possible only due to the hindsight observability that reveals the latent states $x^\ell_{1:H+1}$ for $\ell \in [k-1]$. In general POMDPs, such calculations are not available.

4.2. Regret and sample complexity bounds

We now present the main guarantees for HOP-B in HOMDPs. While we are primarily concerned with sample complexity bounds, HOP-B readily admits a regret bound.

Theorem 4.2. *Let $\mathcal{M}$ be a HOMDP model with X latent states and Y observations. With probability at least $1-\delta$, HOP-B outputs a sequence of policies $\hat{\pi}_1, \dots, \hat{\pi}_K$ such that*

$$\mathrm{Reg}(K) = \widetilde{\mathcal{O}}\left(\sqrt{(XYH^5 + XAH^4)K\iota}\right),$$

where $\iota = \log(2X^2YAKH\delta^{-1})$ and $\widetilde{\mathcal{O}}$ omits lower-order terms in K.

The full bound, including lower order terms, and the proof can be found in Appendix C. A standard online-to-batch conversion reveals that HOP-B learns an ϵ-optimal policy at test time with probability at least $1-\delta$ with sample complexity

$$K = \widetilde{\mathcal{O}}\left(\frac{XYH^5 + XAH^4}{\epsilon^2}\right),$$

omitting log factors and lower-order terms in ϵ. We highlight several conceptual implications of the results.

- The bounds do not have dependence on any *observability parameter*, which typically measures the degree to which one can decode the latent distribution from observations in POMDPs. For instance, guarantees of Liu et al. (2022a) depend polynomially on the inverse of the minimum singular value of $O_\star$ and in fact this dependence is necessary in general POMDPs (see e.g. Theorem 6 in Liu et al., 2022a). This precludes efficient learning in a wide class of partially observed problems (such as vision-based robotics applications with occlusions). By leveraging hindsight observability, our results show that it is possible to circumvent this hardness while still learning a near optimal history-dependent policy for test time.
- The leading terms depend linearly on the number of observations Y and latent states X. Inspecting just the dependence in X and A, the result of Theorem 4.2 can thus be viewed as a natural extension of minimax regret (and sample complexity) results for the MDPs. This is known to be unimprovable in general even for full information MDPs (Dann & Brunskill, 2015; Osband & Van Roy, 2016). However, due to the added complexity of partial observability during deployment, a linear term in Y is also present our bound. Our lower bound in Theorem 5.1 shows that the linear XY dependence is minimax optimal. We remark prior work on POMDPs has yielded large polynomial dependence on X and Y in contrast to our linear dependence here.
- An interesting observation of HOP-B is that the exploration bonus need only happen at the latent state level, rather than needing to explore histories. This suggests that little structure might be needed to learn $O_\star$.[1] We

[1] Indeed, we show in Appendix C.6 that we can incorporate function approximation of $O_\star$ without additional structural conditions beyond realizability as long as the latent model is tabular.

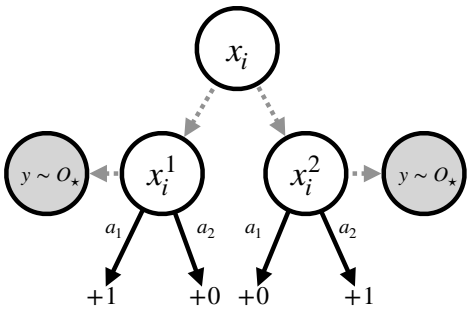

Figure 2: Simplified case of the lower bound construction. The learner starts in x_i and randomly transitions to x_i^1 or x_i^2. The optimal action to achieve reward $+1$ is different depending on which latent state is visited, but the observation y obfuscates sensing of the latent state.

will see in Section 6 that this observation becomes deeper when incorporating function approximation.

The dependence on the horizon H, although still polynomial, is likely suboptimal; however, we conjecture that it can be improved using known tools for MDPs. Since our focus is on the impact of X and Y and understanding the fundamental efficiency gaps between HOMDPs and POMDPs, we leave optimizing these additional factors for future work. As discussed before, HOMDPs are a generalization of latent MDPs with identifiers labeled in hindsight studied by Kwon et al. (2021); Zhou et al. (2022). Ignoring H and specializing our bound to their setting, it matches Kwon et al. (2021). Zhou et al. (2022) is better by a sparsity factor, but is more specialized and not applicable to general HOMDPs.

Proof intuition. HOP-B resembles typical optimistic algorithms for MDPs. However, a naïve analysis by computing confidence intervals on $\hat{T}_k - T_\star$ and $\hat{O}_k - O_\star$ results in an $\mathcal{O}(X^2)$ scaling of sample complexity. We follow the MDP literature in constructing more careful confidence bounds on appropriate value functions instead. This requires some care as value functions are history-dependent in a HOMDP. In particular, we need to control the required exploration as a function of latent state visitations, while reasoning over history-dependent value functions. Combining these ideas carefully, which we present in detail in Appendix C, gives the proof of Theorem 4.2. We note that most POMDP analyses do suffer from the $\mathcal{O}(X^2)$ or worse scaling, as they do not reason via value functions.

5. Limits of Learning in Tabular HOMDPs

We now show that the upper bound of the previous section is optimal in XY for the tabular setting. We present a new information-theoretic lower bound for the tabular HOMDP setting. The proof is in Appendix D.

Theorem 5.1. *Fix $\epsilon \leq 1/64$ and $X, Y \in \mathbb{N}$ such that $Y \geq 6$, $(X+1) \geq 128 \log 2$. For any algorithm $\mathfrak{A}$ producing a policy $\hat{\pi}$ in K episodes of interaction, there exists a HOMDP with the aforementioned cardinalities and $H \asymp \log_2(X)$ and $A = 2$ such that $\mathfrak{A}$ needs*

$$K = \Omega\left(XY/\epsilon^2\right)$$

to guarantee $\mathbb{E}\left[v(\pi^\star) - v(\hat{\pi})\right] \leq \epsilon$, where the expectation is taken over randomness in the data and algorithm.

The lower bound is information-theoretic, meaning that no algorithm can do better than this. The lower bound of Theorem 5.1 matches the XY leading term of the upper bound in Theorem 4.2, suggesting that our algorithm is minimax optimal in X and Y for large K. Recall that existing lower bounds for learning in MDPs necessitate $\Omega\left(XA/\epsilon^2\right)$ episodes of interaction, which accounts for the other leading term in our upper bound (Dann & Brunskill, 2015; Osband & Van Roy, 2016). Since POMDPs are more general than HOMDPs, our lower bound also applies to POMDPs.

Hindsight vs. foresight observability. Our lower bound construction does not distinguish between the latent state x_h being simultaneously revealed along with y_h, or only in hindsight. The key bottleneck is in the construction of the history-dependent policy at test time. Therefore, revealing states simultaneously is no easier statistically than revealing them in hindsight, at least in a minimax sense. This lends credence to framing a broader class of problems such as sim-to-real robotics as HOMDPs by dealing with latent states *after* execution of a policy even though the state is always observable during training. Because one seeks a history-dependent policy in the end, it is just as hard statistically.

Intuition for lower bound construction. We derive the lower bound by constucting a class of hard HOMDP models in the form a binary tree for the latent states, like many hard MDP constructions. At the final layers of the tree is a collection of $\Omega(X)$ subproblems, each of the form of Figure 2. In each subproblem there are 3 latent states. The learner starts in x_i and transitions randomly to either of the child states x_i^1 and x_i^2 with equal probability and must take an optimal action that depends on which one it visits. The difficulty is that $O_\star$ is biased slightly towards half of the observations depending on the latent state. As a result, in order to effectively match the value of the optimal policy, the learner must interact at least $\Omega(Y)$ times with this subproblem. To prove there is linear dependence on XY, we leverage the binary tree described earlier. In short, the learner transitions randomly down the tree until it reaches one of $\Omega(X)$ independent subproblems (decodable from the observations), where it must play optimally. As we remarked earlier, the bottleneck is on the history-dependent policy deployed at test-time, so observing the latent states simultaneously at training does not help in solving this construction.

6. Generalization in HOMDPs

While the tabular setting considered so far is important for building intuition, and tabular latent states are particularly reasonable in many settings, several practical domains of interest necessitate continuous and high-dimensional latent states and observations. In this section, we study HOMDPs where both $\mathcal{X}$ and $\mathcal{Y}$ can be infinitely large, and we employ function approximation for sample-efficient learning. Mirroring prior work in MDPs and more recently POMDPs, this requires structural conditions on the underlying latent state transitions $T_\star$. We now describe one such condition which has been widely studied in the MDP literature, and present an algorithm and sample complexity guarantees.

6.1. Low-rank latent transition dynamics

We initiate this investigation with a well-studied model in the MDP literature: the low-rank MDP (Barreto et al., 2011; Jiang et al., 2017; Agarwal et al., 2020). We study HOMDPs where the underlying latent state MDP is low-rank.

Definition 6.1. A transition function $T_\star$ admits a low-rank decomposition with rank d if there exist vector functions $\phi_\star : \mathcal{X} \times \mathcal{A} \to \mathbb{R}^d$ and $\psi_\star : \mathcal{X} \to \mathbb{R}^d$ such that

$$T_\star(x'|x,a) = \phi_\star(x,a)^\top \psi_\star(x') \quad \forall x, x' \in \mathcal{X}, a \in \mathcal{A}$$

Furthermore, $\phi_\star$ satisfies $\sup_{x,a} \|\phi_\star(x,a)\|_2 \le 1$ and $\psi_\star$ satisfies $\| \int_{x'} \psi_\star(x') \, \mathrm{d}x' \|_2 \le \sqrt{d}$.

Note that in the low-rank setting, we do not assume that the feature embedding $\phi_\star$ is known, unlike the linear MDP setting (Jin et al., 2020b) which crucially leverages this knowledge. We defer details, motivations, and comparisons to the original papers on the matter (Agarwal et al., 2020).

We assume that the learner has access to function classes $\mathcal{T}$ and Θ for approximation of $T_\star$ and $O_\star$, respectively. For simplicity of the analysis, we assume that they are finite but large, and thus we desire a sample complexity which is logarithmic in $|\mathcal{T}|$ and $|\Theta|$, with no explicit dependence on $|\mathcal{X}|$ or $|\mathcal{Y}|$. This formulation also automatically captures the representation learning problem for the latent states (Agarwal et al., 2020). As is standard, we assume that the function classes are proper and satisfy realizability.

Assumption 6.2. $T_\star \in \mathcal{T}$ and $O_\star \in \Theta$. Furthermore, for all $T \in \mathcal{T}$ and $O \in \Theta$, $T(\cdot|x,a) \in \Delta(\mathcal{X})$ and $O(\cdot|x,a) \in \Delta(\mathcal{Y})$ for all $x, x' \in \mathcal{X}$ and $a \in \mathcal{A}$.

The above requires that all candidates in the class can form valid distributions (i.e., they are proper); this can be satisfied by simply discarding those in $\mathcal{T}$ and Θ that are improper.

6.2. The HOP-V algorithm

We now introduce the algorithm Hindsight OPtimism with Version spaces (HOP-V) for function approximation in

Algorithm 2 Hindsight OPtimism with Version spaces (HOP-V)

Input: Transition class $\mathcal{T}$, Emission class Θ.
Set $K' = \lfloor K/H \rfloor$.
Set $\beta_\mathcal{T} = 2\log(K'|\mathcal{T}|/\delta)$.
Set $\beta_\Theta = 2\log(K'|\Theta|/\delta)$.
Initialize $\mathcal{T}_1 = \mathcal{T}$. and $\Theta_1 = \Theta$.
for $k = 1, \dots, K'$ **do**
  `// Optimistic planning`
  Solve

$$\hat{\pi}_k, \hat{O}_k, \hat{T}_k = \operatorname*{arg\,max}_{\pi \in \Pi, T \in \mathcal{T}_k, O \in \Theta_k} v_{\mathcal{M}(T,O)}(\pi)$$

  for $h \in [H]$ **do**
    `// Deploy hist.-dependent policy`
    Construct exploration policy $\tilde{\pi}_k = \hat{\pi}_k \circ_h \mathrm{Unif}(\mathcal{A})$
    Deploy $\tilde{\pi}_k$ and observe trajectory $(y_{1:H}, a_{1:H})$.
    `// Hindsight observation`
    Observe latent states $x_{1:H+1}$.
    Set $y_h^k := y_h$, $a_h^k := a_h$, $x_h^k := x_h$, $\tilde{x}_h^k := x_{h+1}$.
  end for
  Update version spaces with

$$\mathcal{T}_{k+1} = \left\{ T \in \mathcal{T}_k \ : \ \hat{\mathcal{L}}_k^1(T) \ge \max_{T' \in \mathcal{T}_k} \hat{\mathcal{L}}_k^1(T') - \beta_\mathcal{T} \right\}$$

$$\Theta_{k+1} = \left\{ O \in \Theta_k \ : \ \hat{\mathcal{L}}_k^2(O) \ge \max_{O' \in \Theta_k} \hat{\mathcal{L}}_k^2(O') - \beta_\Theta \right\}$$

end for

HOMDP models. HOP-V divides the K rounds into $K' = \lfloor K/H \rfloor$ epochs and maintains version spaces $\mathcal{T}_k$ and Θ_k over the model classes based on the data collected so far for each epoch $k \in [K']$. In epoch k, HOP-V identifies a policy $\hat{\pi}_k$ and models $\hat{T}_k$ and $\hat{O}_k$ by solving an optimistic optimization problem over the version spaces. Here $v_{\mathcal{M}(T,O)}(\pi)$ denotes the value of a policy π in the POMDP model given by transition function T, emission function O and the true reward function r, akin to the original definition in (1). Still within epoch k, for each $h \in [H]$, HOP-V generates an exploration policy from $\hat{\pi}_k$ by taking $\tilde{\pi}_k = \hat{\pi}_k \circ_h \mathrm{Unif}(\mathcal{A})$. The operator $\circ_h$ replaces $\hat{\pi}(\cdot|\tau_h)$ with the uniform distribution $\mathrm{Unif}(\mathcal{A}) \in \Delta(\mathcal{A})$ over the actions for all h-length histories τ_h, but leaves the rest of $\hat{\pi}_k$ unaffected. That is, $\tilde{\pi}_k$ plays $\hat{\pi}_k$ normally up to the hth step and then takes a random action.

HOP-V deploys the exploration policy $\tilde{\pi}_k$ in the environment and records the observation y_h^k and action a_h^k. Then, when the latent states are revealed, it records the latent state x_h^k and the next state $\tilde{x}_h^k$. It repeats this exploration procedure for each $h \in [H]$ in the epoch k to generate $(y_{1:H}^k, a_{1:H}^k, x_{1:H}^k, \tilde{x}_{1:H}^k)$. Based on this new data, it updates the version spaces via maximum likelihood estimation

(MLE) with the following log-likelihood objectives:

$$\hat{\mathcal{L}}_k^1(T) = \sum_{\ell \in [k], h \in [H]} \log T(\tilde{x}_h^\ell | x_h^\ell, a_h^\ell) \quad \text{and}$$
$$\hat{\mathcal{L}}_k^2(O) = \sum_{\ell \in [k], h \in [H]} \log O(y_h^\ell | x_h^\ell).$$

6.3. Sample complexity bound

We now state the performance guarantee for HOP-V.

Theorem 6.3. *Let $\mathcal{M}$ be a HOMDP model with a low-rank transition function $T_\star$ of rank d. Let $\mathcal{T}$ and Θ satisfy Assumption 6.2. Then, with probability at least $1-\delta$, HOP-V outputs a sequence of policies $\hat{\pi}_1, \dots, \hat{\pi}_{K'}$ such that*

$$\mathrm{Reg}(K') = \mathcal{O}\left(\sqrt{AH^4 dK' \log\left(\frac{K'H|\Theta||\mathcal{T}|}{\delta}\right)} \log K'\right)$$

Note that the regret is over the learned $\hat{\pi}_{1:K'}$, not the actually deployed exploration policies. This phenomenon occurs often from one-step exploration (Jiang et al., 2017; Agarwal et al., 2020). However, our focus is the implied PAC guarantee. Again, using a standard online to batch conversion, we get that HOP-V learns an ϵ-optimal policy with probability at least $1-\delta$ in

$$K = \widetilde{\mathcal{O}}\left(\frac{AH^5 d}{\epsilon^2} \log\left(\frac{|\Theta||\mathcal{T}|}{\delta}\right)\right)$$

episodes of interaction. The additional H arises because there are H episodes for each epoch $k \in [K']$ due to the construction and deployment of the exploration policies.

- In contrast to the tabular bound of HOP-B, HOP-V has no dependence on the size of the latent state space X or observation space Y. Instead, generalization using function classes replaces them with complexities of $\mathcal{T}$ and Θ and the rank d of the latent transition. Note that we can readily replace these log-cardinalities with other suitable notions of complexity for infinite function classes.
- For comparison to Theorem 4.2, we can set $d = X$, $\log|\Theta| = \widetilde{\mathcal{O}}(XY)$ and $\log|\mathcal{T}| = \widetilde{\mathcal{O}}(X^2A)$, which results in a suboptimal scaling in X as HOP-V does not use value function-based optimism unlike HOP-B.
- Similar to the tabular setting, the sample complexity also has no dependence on observability parameters, showing that we maintain this advantageous property of HOMDP models in the function approximation setting.
- Observe that we have not made any further structural conditions on $O_\star$ to achieve this result. The structural condition is only on $T_\star$.

Proof intuition. The proof is remarkably simple in contrast to the tabular result. It is a combination of just two components. The first is a simulation lemma (Lemma E.1), which relates the estimation error of the estimated value function to the total variation error of both $\hat{T}_k$ and $\hat{O}_k$. This decomposition allows us the analyze the error almost entirely in terms of the latent state distributions of the exploration policies, rather than their histories. This leads to the second component, which is a standard "one-step-back" analysis that has previously appeared for low-rank MDPs in the fully observable setting (Agarwal et al., 2020). The use of uniform exploration policies at each h is also common in MDP literature to handle the distribution shift between current and historical data. Here the randomness also plays a secondary role of removing all history dependence.

7. Discussion

Motivated by practical applications in partially observable problems, we formulated the problem setting of a Hindsight Observable Markov Decision Process (HOMDP), where the objective is to learn a decision making policy based on partial observations in order to interact with the environment, but the underlying latent states are eventually revealed during the training process. We proposed an algorithm, HOP-B, for finite latent states and observations. We gave sample-complexity upper and lower bounds, showing that HOP-B has no dependence on partial observability parameters and that it is nearly optimal in dependence on XY. We also proposed an algorithm, HOP-V, that allows for function approximation of the transition and emission functions to handle generalization in large or infinite latent state and observation spaces.

There are a number of interesting open directions for future work on hindsight observability. The similarities and compatibility between the HOMDP and standard MDP models make HOMDPs a ripe area for further advancements that leverage our deeper knowledge of MDPs. Natural directions include, for example, model-free RL (Jin et al., 2018; 2020b), offline RL (Xie et al., 2021; Jin et al., 2021b; Zanette et al., 2021), model selection (Lee et al., 2021; 2022; Cutkosky et al., 2021), and computationally efficient representation learning for latent states (Agarwal et al., 2020).

Specific to function approximation, our most general results applied to low-rank latent transition functions for generalization and representation learning; however, MDP literature has had success with more general conditions that restrict the form of the latent state Bellman error (Jiang et al., 2017; Sun et al., 2019; Du et al., 2021; Agarwal & Zhang, 2022). Unfortunately, these conditions have little meaning in HOMDPs and POMDPs because partially observable value functions are history-dependent, not latent state-dependent. It would be interesting in the future to either reconcile these types of conditions or develop new meaningful ones for HOMDPs.

Acknowledgements

Part of this work was done while JNL was a student researcher at Google Research. JNL acknowledges support from the NSF GRFP.

References

Agarwal, A. and Zhang, T. Model-based rl with optimistic posterior sampling: Structural conditions and sample complexity. *arXiv preprint arXiv:2206.07659*, 2022.

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

Auer, P., Jaksch, T., and Ortner, R. Near-optimal regret bounds for reinforcement learning. *Advances in neural information processing systems*, 21, 2008.

Ayer, T., Alagoz, O., and Stout, N. K. Or forum—a pomdp approach to personalize mammography screening decisions. *Operations Research*, 60(5):1019–1034, 2012.

Azar, M. G., Osband, I., and Munos, R. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pp. 263–272. PMLR, 2017.

Azizzadenesheli, K., Lazaric, A., and Anandkumar, A. Reinforcement learning of pomdps using spectral methods. In *Conference on Learning Theory*, pp. 193–256. PMLR, 2016.

Baisero, A. and Amato, C. Unbiased asymmetric actor-critic for partially observable reinforcement learning. *arXiv preprint arXiv:2105.11674*, 2021.

Barreto, A., Precup, D., and Pineau, J. Reinforcement learning using kernel-based stochastic factorization. *Advances in Neural Information Processing Systems*, 24, 2011.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 19–26. JMLR Workshop and Conference Proceedings, 2011.

Cai, Q., Yang, Z., and Wang, Z. Reinforcement learning from partial observation: Linear function approximation with provable sample efficiency. In *International Conference on Machine Learning*, pp. 2485–2522. PMLR, 2022.

Chen, D., Zhou, B., Koltun, V., and Krähenbühl, P. Learning by cheating. In *Conference on Robot Learning*, pp. 66–75. PMLR, 2020.

Chen, F., Bai, Y., and Mei, S. Partially observable rl with b-stability: Unified structural condition and sharp sample-efficient algorithms. *arXiv preprint arXiv:2209.14990*, 2022.

Cutkosky, A., Dann, C., Das, A., Gentile, C., Pacchiano, A., and Purohit, M. Dynamic balancing for model selection in bandits and rl. In *International Conference on Machine Learning*, pp. 2276–2285. PMLR, 2021.

Dann, C. and Brunskill, E. Sample complexity of episodic fixed-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 28, 2015.

Dann, C., Li, L., Wei, W., and Brunskill, E. Policy certificates: Towards accountable reinforcement learning. In *International Conference on Machine Learning*, pp. 1507–1516. PMLR, 2019.

Du, S., Krishnamurthy, A., Jiang, N., Agarwal, A., Dudik, M., and Langford, J. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pp. 1665–1674. PMLR, 2019.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pp. 2826–2836. PMLR, 2021.

Efroni, Y., Jin, C., Krishnamurthy, A., and Miryoosefi, S. Provable reinforcement learning with a short-term memory. *arXiv preprint arXiv:2202.03983*, 2022.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In *International conference on machine learning*, pp. 1126–1135. PMLR, 2017.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Golowich, N., Moitra, A., and Rohatgi, D. Planning in observable pomdps in quasipolynomial time. *arXiv preprint arXiv:2201.04735*, 2022.

Guo, Z. D., Doroudi, S., and Brunskill, E. A pac rl algorithm for episodic pomdps. In *Artificial Intelligence and Statistics*, pp. 510–518. PMLR, 2016.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Jin, C., Kakade, S., Krishnamurthy, A., and Liu, Q. Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 33:18530–18539, 2020a.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*, pp. 2137–2143. PMLR, 2020b.

Jin, C., Liu, Q., and Miryoosefi, S. Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418, 2021a.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021b.

Kamienny, P.-A., Arulkumaran, K., Behbahani, F., Boehmer, W., and Whiteson, S. Privileged information dropout in reinforcement learning. *arXiv preprint arXiv:2005.09220*, 2020.

Krishnamurthy, A., Agarwal, A., and Langford, J. Pac reinforcement learning with rich observations. *Advances in Neural Information Processing Systems*, 29, 2016.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34: 24523–24534, 2021.

Lattimore, T. and Szepesvári, C. *Bandit algorithms*. Cambridge University Press, 2020.

Lee, J., Pacchiano, A., Muthukumar, V., Kong, W., and Brunskill, E. Online model selection for reinforcement learning with function approximation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3340–3348. PMLR, 2021.

Lee, J. N., Tucker, G., Nachum, O., Dai, B., and Brunskill, E. Oracle inequalities for model selection in offline reinforcement learning. *arXiv preprint arXiv:2211.02016*, 2022.

Liu, E. Z., Raghunathan, A., Liang, P., and Finn, C. Decoupling exploration and exploitation for meta-reinforcement learning without sacrifices. In *International conference on machine learning*, pp. 6925–6935. PMLR, 2021.

Liu, Q., Chung, A., Szepesvári, C., and Jin, C. When is partially observable reinforcement learning not scary? *arXiv preprint arXiv:2204.08967*, 2022a.

Liu, Q., Netrapalli, P., Szepesvari, C., and Jin, C. Optimistic mle–a generic model-based algorithm for partially observable sequential decision making. *arXiv preprint arXiv:2209.14997*, 2022b.

Massart, P. *Concentration inequalities and model selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003*. Springer, 2007.

Osband, I. and Van Roy, B. On lower bounds for regret in reinforcement learning. *arXiv preprint arXiv:1608.02732*, 2016.

Pan, Y., Cheng, C.-A., Saigol, K., Lee, K., Yan, X., Theodorou, E., and Boots, B. Agile autonomous driving using end-to-end deep imitation learning. *arXiv preprint arXiv:1709.07174*, 2017.

Papadimitriou, C. H. and Tsitsiklis, J. N. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

Pinto, L., Andrychowicz, M., Welinder, P., Zaremba, W., and Abbeel, P. Asymmetric actor critic for image-based robot learning. *arXiv preprint arXiv:1710.06542*, 2017.

Regatti, J. R., Deshmukh, A. A., Cheng, F., Jung, Y. H., Gupta, A., and Dogan, U. Offline rl with resource constrained online deployment. *arXiv preprint arXiv:2110.03165*, 2021.

Ross, S., Gordon, G., and Bagnell, D. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pp. 627–635. JMLR Workshop and Conference Proceedings, 2011.

Shani, G., Heckerman, D., Brafman, R. I., and Boutilier, C. An mdp-based recommender system. *Journal of Machine Learning Research*, 6(9), 2005.

Sinclair, S. R., Frujeri, F., Cheng, C.-A., and Swaminathan, A. Hindsight learning for mdps with exogenous inputs. *arXiv preprint arXiv:2207.06272*, 2022.

Smallwood, R. D. and Sondik, E. J. The optimal control of partially observable markov processes over a finite horizon. *Operations research*, 21(5):1071–1088, 1973.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pp. 2898–2933. PMLR, 2019.

Thrun, S. Probabilistic algorithms in robotics. *Ai Magazine*, 21(4):93–93, 2000.

Uehara, M., Sekhari, A., Lee, J. D., Kallus, N., and Sun, W. Computationally efficient pac rl in pomdps with latent determinism and conditional embeddings. *arXiv preprint arXiv:2206.12081*, 2022.

Wang, J. X., Kurth-Nelson, Z., Tirumala, D., Soyer, H., Leibo, J. Z., Munos, R., Blundell, C., Kumaran, D., and Botvinick, M. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

Warrington, A., Lavington, J. W., Scibior, A., Schmidt, M., and Wood, F. Robust asymmetric learning in pomdps. In *International Conference on Machine Learning*, pp. 11013–11023. PMLR, 2021.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

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

Zanette, A. and Brunskill, E. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *International Conference on Machine Learning*, pp. 7304–7312. PMLR, 2019.

Zanette, A., Wainwright, M. J., and Brunskill, E. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–13640, 2021.

Zhan, W., Uehara, M., Sun, W., and Lee, J. D. Pac reinforcement learning for predictive state representations. *arXiv preprint arXiv:2207.05738*, 2022.

Zhong, H., Xiong, W., Zheng, S., Wang, L., Wang, Z., Yang, Z., and Zhang, T. A posterior sampling framework for interactive decision making. *arXiv preprint arXiv:2211.01962*, 2022.

Zhou, R., Wang, R., and Du, S. S. Horizon-free reinforcement learning for latent markov decision processes. *arXiv preprint arXiv:2210.11604*, 2022.

Contents of Appendix

Image sources: The globe in Figure 1 is taken from https://commons.wikimedia.org/wiki/File:Ambox_globe_Americas.svg.

A. Additional Motivating Applications

To further establish hindsight observability, we describe in more detail a number of motivating applications. The diversity of the problem settings highlights the generality of the HOMDP model.

- **Sim-to-real robotics.** Sim-to-real (simulation to reality) is a well-studied paradigm in robotics in which one trains a robot using RL in a simulator with the end goal of deploying the robot in the real world at test time. The test time policy often uses high dimensional, partial observations such as images from a camera (which are susceptible to noise and occlusions). However, during training, the simulator grants access to the underlying state (object positions, poses, etc.), information that is not available at test time but might dramatically increase sample efficiency. Empirically, leveraging this train-time information has led to improved sample complexity (Pinto et al., 2017; Chen et al., 2020).

- **High-frequency control.** In a related control setting, latency and computational bottlenecks (e.g. processing lidar data, depth images, etc.) can obscure and delay observation of the true state even though control inputs must be made quickly. It is common in such settings to instead use simpler observations, such as images from a standard camera. Once the system finally observes or processes the true states, the states can be incorporated in the training procedure. Pan et al. (2017) explored this direction, successfully training an autonomous rally car via distillation for high-speed driving.

- **Meta-learning and latent MDPs.** In meta-learning for RL (Wang et al., 2016; Finn et al., 2017), an agent leverages past rollouts on different MDP tasks sampled from a fixed distribution, where the reward and transition functions differ from task to task. The training tasks are often labeled or can be inferred in hindsight (Liu et al., 2021). At test time, the task is unknown and thus the agent must adapt using only observations to infer and maximize reward for the test time task. This can be modeled as a HOMDP, where the latent states are the MDP states concatenated with the task context and the observations are the MDP states. Meta-learning with finitely many tasks can also be viewed as a latent MDP (Kwon et al., 2021; Zhou et al., 2022).

- **Scheduling.** Following Sinclair et al. (2022), consider a data center, which aims to allocate submitted jobs to servers efficiently. The arrival times of the jobs and their total lengths are unknown. However, once a job has completed, both its arrival time and length are known in hindsight. The latent states of the equivalent HOMDP are the observations concatenated with the arrival times and lengths of all jobs.

- **Online imitation learning** In online imitation learning (Ross et al., 2011), an agent interacts with the MDP over K rounds, executing actions under its learned policy. After a round, it retroactively queries an expert for optimal actions in the visited states to then update the policy. This can be viewed as a special case of our setting where each of our latent states is a visited state concatenated with the expert's optimal action for that state, which is only retroactively observed.

- **Screening for diseases.** Medical screenings are procedures designed to help detect and monitor diseases in patients, usually in early stages. Screenings are typically dependent on patient characteristics as well as tests, which may not always be accurate and may have undesirable effects. Ayer et al. (2012) frame the problem of screening for breast cancer as a POMDP, where the state is a patient's true condition (progression of the disease), and the observations are the outcomes of tests such as a mammogram. Depending on the actions taken in response to the observed tests, the patient may eventually undergo a perfect test revealing the latent state (such as a biopsy). Revelation of the state can then be used to more effectively guide observation-based policies in the future by allowing association between the outcomes of prior tests and the true condition.

B. Value Functions and Alpha Vector Representations

The proofs in this paper crucially rely on the α-vector representation of value functions for POMDPs (Smallwood & Sondik, 1973). Since these techniques and intuitions are not common in reinforcement learning theory literature[2], we now give an introductory treatment of this topic with all the tools necessary for our proofs.

[2]Exceptions include Kwon et al. (2021); Zhou et al. (2022) who used α-vectors in the latent MDP model (single unobserved context).

Since we work with history-dependent policies, the value V^π and action-value functions Q^π of a policy π are history-dependent as well. The history of observations and actions τ_h gives rise to a posterior distribution $b_h(\cdot) = P(x_h = \cdot|\tau_h)$ over the latent states.[3] We could define value functions simply as a function of τ_h but it will be convenient to make the posterior b_n explicit and write value functions as a function of both x_h and b_h:

$$Q_h^\pi(b_h, \tau_h, a_h) = \mathbb{E}_{x_h \sim b_h}\left[\mathbb{E}\left[\sum_{h'=h}^{H} r(x_h, a_h) \mid \tau_h, x_h, a_h\right]\right]$$
$$V_h^\pi(b_h, \tau_h) = \sum_{a_h} \pi(a_h|\tau_h) Q_h^\pi(b_h, \tau_h, a_h).$$

While b_h determines the latent state, τ_h affects the actions taken by the policy.

The α-vectors of a POMDP act as a kind of representation for the value functions of possible policies on the POMDP. This allows us to represent value functions as linear functions of belief vectors. Note that it is not obvious how to write either of the value function V or Q for POMDPs in a concise, recursive form in the same way that we do for MDP value functions. The α-vector representation provides an alternative solution where we can represent the value functions as linear functions with the α-vectors and then have a recursive definition of the α-vectors.

Consider a POMDP with transtion function T and emission function O. Consider the last timestep H and a belief distribution $b \in \Delta(\mathcal{X})$ over the latent state:

$$Q_H^\pi(b, \tau, a) = \sum_x b(x) r(x, a) \tag{4}$$
$$V_H^\pi(b, \tau) = \sum_{x,a} b(x) \pi(a|\tau) r(x, a). \tag{5}$$

These clearly have simple linear representations as functions of the belief vector:

$$Q_H^\pi(b, \tau, a) = b^\top \alpha_{H,\tau}^\pi(\cdot, a) \tag{6}$$
$$V_H^\pi(b, \tau) = b^\top \alpha_{H,\tau}^\pi(\cdot), \tag{7}$$

where τ is a partial history of appropriate length, $\alpha_{H,\tau}^\pi(\cdot) \in \mathbb{R}^X$ and $\alpha_{H,\tau}^\pi(\cdot, \cdot) \in \mathbb{R}^{X \times A}$ are given by

$$\alpha_{H,\tau}^\pi(x, a) = r(x, a) \tag{8}$$
$$\alpha_{H,\tau}^\pi(x) = \sum_a \pi(a|\tau) \alpha_{H,\tau}^\pi(x, a). \tag{9}$$

Let us now assume inductively that $V_{h+1}^\pi(x)$ and Q_{h+1}^π have the following representations:

$$Q_{h+1}^\pi(b, \tau, a) = b^\top \alpha_{h+1,\tau}^\pi(\cdot, a) \tag{10}$$
$$V_{h+1}^\pi(b, \tau) = b^\top \alpha_{h+1,\tau}^\pi. \tag{11}$$

From the definition of Q,

$$Q_h^\pi(b, \tau, a) = b^\top r(\cdot, a) + \sum_{y'} P(y'|\tau, a) V_{h+1}^\pi(b', \tau') \tag{12}$$

where we denote τ' as the history τ concatenated with the new action a and observation y' and where b' (which is dependent on y' and a) is updated belief vector starting from b given action a and the next observation y'.

$$b'(x') := \frac{O(y'|x') \sum_x b(x) T(x'|x, a)}{\sum_{x''} O(y'|x'') \sum_x b(x) T(x''|x, a)}. \tag{13}$$

[3] The posterior depends on the initial latent state distribution ρ but we omit a notational references to it for clarity.

The action-value function can then be rewritten as

$$Q_h^\pi(b,\tau,a) = b^\top r(\cdot,a) + \sum_{y'} P(y'|\tau,a) \sum_{x'} b'(x') \alpha_{h+1,\tau'}^\pi(x') \tag{14}$$

$$= b^\top r(\cdot,a) + \sum_{x,x',y'} b(x) T(x'|x,a) O(y'|x') \alpha_{h+1,\tau'}^\pi(x') \tag{15}$$

$$= b^\top r(\cdot,a) + b^\top \gamma_{h,\tau}^\pi(\cdot,a), \tag{16}$$

where we define

$$\gamma_{h,\tau}^\pi(x,a) = \sum_{x',y'} T(x'|x,a) O(y'|x') \alpha_{h+1,\tau'}^\pi(x') \tag{17}$$

$$\gamma_{h,\tau}^\pi(x) = \sum_{a,x',y'} \pi_h(a|\tau) T(x'|x,a) O(y'|x') \alpha_{h+1,\tau'}^\pi(x'). \tag{18}$$

Then, define

$$\alpha_{h,\tau}^\pi(\cdot,a) = r(\cdot,a) + \gamma_{h,\tau}^\pi(\cdot,a) \tag{19}$$

$$\alpha_{h,\tau}^\pi(\cdot) = \sum_a \pi_h(a|b) r(\cdot,a) + \gamma_{h,\tau}^\pi(\cdot). \tag{20}$$

This enables

$$Q_h^\pi(b,\tau,a) = b^\top \alpha_{h,\tau}^\pi(\cdot,a) \tag{21}$$

$$V_h^\pi(b,\tau) = b^\top \alpha_{h,\tau}^\pi(\cdot). \tag{22}$$

We may repeat this recursion until the end $h = 1$.

This culminates in the following α-vector proposition.

Proposition B.1 (α-vector representation)**.** *Let π be a fixed history-dependent policy. Let b_h denote the belief vector (posterior distribution over $\mathcal{X}$) given the history $\tau_h = (y_{1:h}, a_{1:h-1})$. Then, there exist vectors $\alpha_{h,\tau_h}^\pi(\cdot) \in \mathbb{R}^X$ and $\alpha_{h,\tau_h}^\pi(\cdot,a) \in \mathbb{R}^X$ such that, for all (h,x,a,τ_h), the following equations hold:*

$$V_h^\pi(b_h,\tau_h) = b_h^\top \alpha_{h,\tau_h}^\pi(\cdot) \tag{23}$$

$$Q_h^\pi(b_h,\tau_h) = b_h^\top \alpha_{h,\tau_h}^\pi(\cdot,a) \tag{24}$$

$$\alpha_{h,\tau_h}^\pi(x,a) = r(x,a) + \sum_{x',y'} T(x'|x,a) O(y'|x') \alpha_{h+1,\tau'}^\pi(x') \tag{25}$$

$$\alpha_{h,\tau_h}^\pi(x) = \sum_a \pi(a|\tau_h) \alpha_{h,\tau_h}^\pi(x,a), \tag{26}$$

where τ' is the concatenation of τ with observation y' and action a, and $\alpha_{H+1} = 0$. Furthermore, $\max\left\{\alpha_{h,\tau_h}^\pi(x,a), \alpha_{h,\tau_h}^\pi(x)\right\} \le G(H-h+1)$ if $r(x,a) \in [0,G]$ for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$ and $h \in [H]$.

Proof. We have already proved the equations by induction. It remains to show that the values we constructed satisfy the last statement, the bound. We will focus on $\alpha_{h,\tau}^\pi(x,a)$ since it is clear that if the bound is satisfied for this one, then it is satisfied for $\alpha_{h,\tau}^\pi(x)$. Using proof by induction, we have the base case $\alpha_{H,\tau}^\pi(x,a) \le \max_{x,a} r(x,a) \in [0,G]$. Then,

$$\alpha_{h,\tau}^\pi(x,a) = r(x,a) + \sum_{x',y'} T(x'|x,a) O(y'|x') \alpha_{h+1,\tau'}^\pi(x') \tag{27}$$

$$\le r(x,a) + \sum_{x',y'} T(x'|x,a) O(y'|x') G\left(H - (h+1) + 1\right) \tag{28}$$

$$= r(x,a) + G(H-h) \tag{29}$$

$$\le G(H-h+1). \tag{30}$$

This concludes the proof.

□

C. Full Statement and Proof of Theorem 4.2

Having established the notation and important concepts of the α-vectors, we now proceed to the full statement (including all lower order terms) and the proof of Theorem 4.2, which will immediately put these concepts to use.

C.1. Full statement of result of Theorem 4.2

We let $a \lesssim b$ mean that $a \leq cb$ where c is a problem-independent constant.

Theorem C.1. *Let $\mathcal{M}$ be a HOMDP model with X latent states and Y observations. With probability at least $1-5\delta$, HOP-B outputs a sequence of policies $\hat{\pi}_1, \ldots, \hat{\pi}_K$ such that*

$$\sum_{k\in[K]} v(\pi^\star) - v(\hat{\pi}_k) \lesssim \underbrace{\sqrt{H^5 K \log(2/\delta)}}_{\textit{Azuma-Hoeffding}} + \underbrace{\sqrt{YXH^5K\iota}}_{\textit{Emission error}} + \underbrace{\sqrt{XAH^4K\iota} + H^4X^2A\iota(1+\log(K))}_{\textit{Transition error}} \tag{31}$$

$$+ \underbrace{H^3X\sqrt{Y\iota} + HXA\sqrt{H^3\iota}}_{\textit{Residual pigeonhole error}}, \tag{32}$$

where $\iota = \log(2X^2YAKH/\delta)$.

C.2. High-probability events

We begin by defining several events that we show occur with high probability. For the optimal policy $\pi^\star$, there exist vectors $\alpha^{\pi^\star}_{h,\tau_h} \in \mathbb{R}^X$, indexed by latent states, such that $V^{\pi^\star}_h(b_h, \tau_h) = b_h^\top \alpha^{\pi^\star}_{h,\tau_h}$ (see Proposition B.1). As such, we define the following event that bounds the deviation on $T_\star$ by leveraging $\alpha^{\pi^\star}$ in a similar manner to how the optimal value functions are leveraged in improved MDP analyses such as Azar et al. (2017). Define

$$\mathcal{E}_T = \left\{ \forall k, h, x, a, \tau_h, \sum_{y',x'} O_\star(y'|x') \left(T_\star(x'|x,a) - \hat{T}_k(x'|x,a)\right) \alpha^{\pi_\star}_{h+1,\tau'_h}(x') \leq \sqrt{\frac{C_T H^3 \log(YXAHK/\delta)}{n_k(x,a)}} \right\}, \tag{33}$$

where $C_T = 4$. For $h = H$, the left side is just zero. Recall that τ'_h denotes the concatenated partial history of τ_h with (y', a) (i.e. the "next-step" partial history. The purpose of this event is to avoid resorting to a total variation bound which necessitates immediate dependence on X. By instead bounding quantities solely in terms of the α-vector of the optimal policy, $\alpha^{\pi^\star}_{h+1,\tau'_h}(x')$, we can still produce valid bonuses. Such tricks are often used to get the best MDP regret bounds (Azar et al., 2017). However, in contrast to the MDP style analyses, the above event must hold across all possible observable histories τ_h which leads to additional polynomial dependence on H (due to there being exponentially many histories in H and a union bound over these histories) and polylogarithmic dependence on Y. However, using this approach will save a X factor in the sample complexity bound.

We also consider the following alternative version of $\mathcal{E}_T$. Let $c \geq 1$ be a constant, potentially dependent on H. Define

$$\mathcal{E}^c_T = \left\{ \forall k, x, a, x',\ \hat{T}_k(x'|x,a) - T_\star(x'|x,a) = \frac{T_\star(x'|x,a)}{2c} + \frac{2c\log(X^2AKH/\delta)}{n_k(x,a)} \right\}. \tag{34}$$

To handle estimation of the emission matrix, we will use a more conventional event based on the total variation difference of $O_\star$ and the estimated quantity $\hat{O}_k$:

$$\mathcal{E}_O = \left\{ \forall k \in [K], x \in \mathcal{X},\ \|O_\star(\cdot|x) - \hat{O}_k(\cdot|x)\|_1 \leq \sqrt{\frac{C_O Y \log(YXKH/\delta)}{n_k(x)}} \right\}, \tag{35}$$

where $C_O = 8$.

Lemma C.2. $P(\mathcal{E}_T) \geq 1-\delta$.

Proof. For each x, a we will assume that KH independent transitions are preemptively sampled from $T(\cdot|x,a)$ and revealed in order to the learner with each visit to (x,a), since this is distributionally identical to the interface the learner encounters. Let $\hat{T}_{(n)}(\cdot|x,a)$ denote the empirical distribution estimated with the first $n \in [KH]$ samples. We can then apply Hoeffding's

inequality to a specific sum of independent variables. Consider fixed $n \in [KH], x, a$ and an arbitrary function $g : \mathcal{X} \to \mathbb{R}$ such that $\|g\|_\infty \leq G$ uniformly for some $G \geq 0$. Here, we will use $x'_{(i)}$ to denote the realizations of the preemptive samples for $i \in [n]$. Then, with probability at least $1 - \delta$,

$$\sum_{x'} g(x') \left(\hat{T}_{(n)}(x'|x,a) - T(x'|x,a) \right) = \frac{1}{n} \sum_{i \in [n]} \left(\sum_{x'} g(x') \left(\mathbf{1}\{x'_{(i)} = x'\} - T_\star(x'|x,a) \right) \right). \tag{36}$$

Since this is a sum of independent random variables bounded within $[-G, G]$, Hoeffding's inequality implies that

$$\sum_{x'} g(x') \left(\hat{T}_{(n)}(x'|x,a) - T_\star(x'|x,a) \right) \leq 2G \sqrt{\frac{\log(1/\delta)}{2n}} \tag{37}$$

with probability at least $1 - \delta$. Then, we can simply choose $g(x') = \sum_{y'} O_\star(y'|x') \alpha^{\pi_*}_{h+1,\tau'_h}(x')$, which is fixed and has $|g(x')| \leq (H - h)$ via the bound on the size of the α-vectors due to Proposition B.1. Taking the union bound over all $n \in [KH], h \in [H], x \in \mathcal{X}, a \in \mathcal{A}, \tau_h \in \mathcal{Y}^h \times \mathcal{A}^{h-1}$, we have

$$2(H-h) \sqrt{\frac{\log\left(Y^h X A^{h-1} H^2 K/\delta\right)}{2n}} \leq \sqrt{C_T H^3 \log\left(YXAHK/\delta\right)},$$

which gives the result with constant $C_T = 4$. □

Lemma C.3. $P(\mathcal{E}_T^c) \geq 1 - \delta$.

Proof. We again use the distributionally equivalent notation from the prior proof and the same notation for n. By Bernstein's inequality (Lemma F.3), with probability at least $1 - \delta$,

$$\hat{T}_{(n)}(x'|x,a) - T_\star(x'|x,a) \leq \frac{T_\star(x'|x,a)}{2c} + \frac{2c \log(1/\delta)}{n}. \tag{38}$$

Taking the union bound over all $n \in [KH]$, $x, x' \in \mathcal{X}$, and $a \in \mathcal{A}$ gives the result. □

$\mathcal{E}_T^c$ immediately implies the following error bound.

Corollary C.4. *Let $g : \mathcal{X} \to [-G, G]$ be a bounded function for $G \geq 0$. Suppose that $\mathcal{E}_T^c$ holds. Then, for all x, a, k,*

$$\sum_{x'} \left(\hat{T}_k(x'|x,a) - T_\star(x'|x,a) \right) g(x') \leq \frac{1}{2c} \sum_{x'} T_\star(x'|x,a) g(x') + \frac{2cGX \log(X^2 AKH/\delta)}{n_k(x,a)}. \tag{39}$$

Proof. The proof is immediate by rearranging the definition of $\mathcal{E}_T^c$. □

Lemma C.5. $P(\mathcal{E}_O) \geq 1 - \delta$.

Proof. A similar approach via Bernstein's inequality (Lemma F.3) guarantees the following for the analogously defined $\hat{O}_{(n)}(y|x)$ for $n \in [KH]$. With probability at least $1 - \delta$, for all x, y, n,

$$|\hat{O}_{(n)}(y|x) - O_\star(y|x)| \leq \sqrt{\frac{2O_\star(y|x)\iota}{n}} + \frac{\iota}{3n}, \tag{40}$$

where $\iota = \log(2YXKH/\delta)$. Therefore,

$$\frac{1}{2} \|\hat{O}_{(n)}(\cdot|x) - O_\star(\cdot|x)\|_1 \leq \frac{Y\iota}{6n} + \sum_y \sqrt{\frac{O_\star(y|x)\iota}{2n}} \tag{41}$$

$$\leq \frac{Y\iota}{2n} + \sqrt{\frac{Y\iota}{2n}} \tag{42}$$

$$\leq \sqrt{\frac{2Y\iota}{n}}, \tag{43}$$

where the last line follows from the fact that $\frac{1}{2}\|\hat{O}_{(n)}(\cdot|x) - O_\star(\cdot|x)\|_1 \leq 1$ always. So if the upper bound of the right side is at most 1 then $\frac{Y\iota}{2n}$ is at most 1. Therefore, the desired bound holds with $C_O = 8$. □

C.3. Optimism via reward bonuses

We let $\hat{\alpha}_k$ and α denote the α-vectors under the learned model $\widehat{\mathcal{M}}_k$, which uses the transition function $\hat{T}_k$ and emission function $\hat{O}_k$ and bonus reward function $\hat{r}_k$, and the true model $\mathcal{M}$, respectively.

Lemma C.6. *Suppose that $\mathcal{E}_T$ and $\mathcal{E}_O$ hold. For all k, h, τ_h, x, it holds that $\hat{\alpha}^{\pi^\star}_{k,h,\tau_h}(x) \geq \alpha^{\pi^\star}_{h,\tau_h}(x) + H\epsilon_k(x)$ for all $h \in [H]$.*

Proof. The proof is by induction on h. Let k be fixed so we can drop the subscript notation for it. Observe that we clearly have the base case via Proposition B.1:

$$\begin{aligned}
\hat{\alpha}^{\pi^\star}_{H,\tau_H}(x,a) &= \hat{r}(x,a) && (44)\\
&= r(x,a) + H\epsilon(x) + \epsilon(x,a) && (45)\\
&= \alpha^{\pi^\star}_{H,\tau_H}(x,a) + H\epsilon(x) + \epsilon(x,a) && (46)\\
&\geq \alpha^{\pi^\star}_{H,\tau_H}(x,a) + H\epsilon(x). && (47)
\end{aligned}$$

Fix $h \in [H-1]$. Recall the definition of τ'_h as the "next-step" partial history. Assume that $\hat{\alpha}^{\pi^\star}_{h+1,\tau'_h}(x) \geq \hat{\alpha}^{\pi^\star}_{h+1,\tau'_h}(x) + H\epsilon(x)$. Then,

$$\begin{aligned}
\alpha^{\pi^\star}_{h,\tau_h}(x,a) &= r(x,a) + \sum_{x',y'} T_\star(x'|x,a) O_\star(y'|x') \alpha^{\pi^\star}_{h+1,\tau'_h}(x') && (48)\\
&= r(x,a) + \sum_{x',y'} O_\star(y'|x') \left(T_\star(x'|x,a) - \hat{T}_\star(x'|x,a)\right) \alpha^{\pi^\star}_{h+1,\tau'_h}(x') && (49)\\
&\quad + \sum_{x',y'} \hat{T}(x'|x,a) \left(O_\star(y'|x') - \hat{O}(y'|x')\right) \alpha^{\pi^\star}_{h+1,\tau'_h}(x') && (50)\\
&\quad + \sum_{x',y'} \hat{T}(x'|x,a) \hat{O}(y'|x') \alpha^{\pi^\star}_{h+1,\tau'_h}(x'). && (51)
\end{aligned}$$

The first summation is bounded using $\mathcal{E}_T$:

$$\begin{aligned}
\sum_{x',y'} O_\star(y'|x') \left(T_\star(x'|x,a) - \hat{T}(x'|x,a)\right) \alpha^{\pi^\star}_{h+1,\tau'_h}(x') &\leq \sqrt{\frac{C_T H^3 \log(YXAHK/\delta)}{n(x,a)}} && (52)\\
&\leq \epsilon(x,a). && (53)
\end{aligned}$$

The second summation can be bounded in terms of the total variation distance for the emission matrices along with $\mathcal{E}_O$:

$$\begin{aligned}
\sum_{x',y'} \hat{T}(x'|x,a) \left(O_\star(y'|x') - \hat{O}(y'|x')\right) \alpha^{\pi^\star}_{h+1,\tau'_h}(x') &\leq \sum_{x'} (H-h) \hat{T}(x'|x,a) \|O_\star(\cdot|x') - \hat{O}(\cdot|x')\|_1 && (54)\\
&\leq \sum_{x'} (H-h) \hat{T}(x'|x,a) \epsilon(x'), && (55)
\end{aligned}$$

which also uses Proposition B.1 to bound the magnitude of $\alpha^{\pi^\star}$. Finally, for the third summation, we can use the inductive hypothesis to get

$$\sum_{x',y'} \hat{T}(x'|x,a) \hat{O}(y'|x') \alpha^{\pi^\star}_{h+1,\tau'_h}(x') \leq \sum_{x',y'} \hat{T}(x'|x,a) \hat{O}(y'|x') \left(\hat{\alpha}^{\pi^\star}_{h+1,\tau'_h}(x') - H\epsilon(x')\right). \quad (56)$$

Combining these three individual bounds, we have

$$\begin{aligned}
\alpha^{\pi^\star}_{h,\tau_h}(x,a) &\leq r(x,a) + \epsilon(x,a) + \sum_{x'} (H-h) \hat{T}(x'|x,a) \epsilon(x') && (57)\\
&\quad + \sum_{x',y'} \hat{T}(x'|x,a) \hat{O}(y'|x') \left(\hat{\alpha}^{\pi^\star}_{h+1,\tau'_h}(x') - H\epsilon(x')\right) && (58)\\
&\leq r(x,a) + \epsilon(x,a) + \sum_{x',y'} \hat{T}(x'|x,a) \hat{O}(y'|x') \hat{\alpha}^{\pi^\star}_{h+1,\tau'_h}(x') && (59)\\
&= \hat{\alpha}^{\pi^\star}_{h,\tau_h}(x,a) - H\epsilon(x), && (60)
\end{aligned}$$

where we have added and subtracted $H\epsilon(x)$ and used the definition of $\hat{\alpha}^{\pi^\star}_{h,\tau_h}$ (Propostion B.1) in the last step. Applying this inductively gives the result.

□

Lemma C.7. *For all history-dependent π and all k, h, τ_h, it holds that $\|\hat{\alpha}^{\pi}_{k,h,\tau_h}\|_\infty \le 5H(H-h+1)$.*

Proof. As before, we assume k is fixed and drop subscript notation for it. Note that we have $\hat{r}(x,a) = r(x,a) + H\epsilon(x) + \epsilon(x,a)$. Furthermore,

$$\epsilon(x) \le 2 \tag{61}$$
$$\epsilon(x,a) \le 2H. \tag{62}$$

Therefore, $\hat{r}(x,a) \in [0, 5H]$. Proposition B.1 implies the result. □

C.4. Proof of the theorem

Define the intersection of the above events as $\mathcal{E} = \mathcal{E}_T \cap \mathcal{E}_O \cap \mathcal{E}_T^c \cap \mathcal{E}_T^1$ for some fixed c to be determined later. Note that $P(\mathcal{E}) \ge 1 - 4\delta$ by the union bound. For the remainder of the proof, we shall assume that these four hold simultaneously. Recall also that we define the errors as

$$\epsilon_k(x,a) := \min\left\{2H,\ \sqrt{\frac{C_T H^3 \log(YXAHK/\delta)}{n_k(x,a)}}\right\}, \tag{63}$$

$$\epsilon_k(x) := \min\left\{2,\ \sqrt{\frac{C_O Y \log(YXKH/\delta)}{n_k(x)}}\right\}. \tag{64}$$

We define an additional error term that is not used in the algorithm, only the analysis:

$$\tilde{\epsilon}_k(c,x,a) := \min\left\{2, \frac{2cX\log(X^2AKH/\delta)}{n_k(x,a)}\right\}. \tag{65}$$

We begin by analyzing a fixed round k and will temporarily drop subscripts denoting k. We will use $\hat{\mathbb{E}}$ and $\hat{P}$ to denote expectation and probabilities under the learned model $\widehat{\mathcal{M}}$ during this round, as defined in the previous subsection. We let $\hat{v}(\pi)$ denote the average value of the policy π under $\widehat{\mathcal{M}}$, akin to the true value $v(\pi)$. Then, using the α-vector definition of the value functions,

$$\hat{v}(\pi^\star) - v(\pi^\star) = \sum_{x,y} \rho(x)\hat{O}(y|x)\hat{\alpha}^{\pi^\star}_{1,\tau_1}(x) - \sum_{x,y}\rho(x)O_\star(y|x)\alpha^{\pi^\star}_{1,\tau_1}(x) \tag{66}$$

$$\ge -H\sum_x \rho(x)\epsilon(x) + \sum_{x,y}\rho(x)\hat{O}(y|x)\left(\hat{\alpha}^{\pi^\star}_{1,\tau_1}(x) - \alpha^{\pi^\star}_{1,\tau_1}(x)\right) \tag{67}$$

$$\ge 0, \tag{68}$$

where the last inequality follows from the optimism bound in Lemma C.6. Therefore, by $\mathcal{E}_O$ and Lemma C.7,

$$v(\pi^\star) - v(\hat{\pi}) \le \hat{v}(\hat{\pi}) - v(\hat{\pi}) \tag{69}$$

$$= \sum_{x,y}\rho(x)\hat{O}(y|x)\hat{\alpha}^{\hat{\pi}}_{1,\tau_1}(x) - \sum_{x,y}\rho(x)O_\star(y|x)\alpha^{\hat{\pi}}_{1,\tau_1}(x) \tag{70}$$

$$\le 5H^2\mathbb{E}_{\hat{\pi}}\left[\epsilon(x_1)\right] + \mathbb{E}_{\hat{\pi}}\left[\hat{\alpha}^{\hat{\pi}}_{1,\tau_1}(x_1) - \alpha^{\hat{\pi}}_{1,\tau_1}(x_1)\right] \tag{71}$$

The crux of the proof lies in the following lemma which recursively bounds the expected differences of the α-vectors under $\hat{\pi}$. For convenience, let us set $C := 21$.

Lemma C.8. *Let $h \in [H-1]$ be fixed. Then, under the aforementioned events, it holds that*

$$\mathbb{E}_{\hat{\pi}}\left[\hat{\alpha}^{\hat{\pi}}_{h,\tau_h}(x_h) - \alpha^{\hat{\pi}}_{h,\tau_h}(x_h)\right] \le \mathbb{E}_{\hat{\pi}}\left[H\epsilon(x_h) + 2\epsilon(x_h,a_h) + CH^2\tilde{\epsilon}(c,x_h,a_h)\right] \tag{72}$$

$$+ \left(1+\frac{1}{2c}\right)\mathbb{E}_{\hat{\pi}}\left[\hat{\alpha}^{\hat{\pi}}_{h+1,\tau_{h+1}}(x_{h+1}) - \alpha^{\hat{\pi}}_{h+1,\tau_{h+1}}(x_{h+1}) + 11H^2\epsilon(x_{h+1})\right]. \tag{73}$$

Furthermore,

$$\mathbb{E}_{\hat{\pi}}\left[\hat{\alpha}^{\hat{\pi}}_{H,\tau_H}(x_H) - \alpha^{\hat{\pi}}_{H,\tau_H}(x_H)\right] \le \mathbb{E}_{\hat{\pi}}\left[H\epsilon(x_H) + \epsilon(x_H,a_H)\right]. \tag{74}$$

Lemma C.8 is proved in Appendix C.5.

Now we can apply the bound from Lemma C.8 recursively to get the following bound on the value difference under the true and learned models:

$$\hat{v}(\hat{\pi}) - v(\hat{\pi}) \le \left(1+\frac{1}{2c}\right)^H \mathbb{E}_{\hat{\pi}}\left[\sum_h 12H^2\epsilon(x_h) + 2\epsilon(x_h,a_h) + CH^2\tilde{\epsilon}(c,x_h,a_h)\right]. \tag{75}$$

Therefore, we can choose $c = {}^H\!/_2$ to get

$$\hat{v}(\hat{\pi}) - v(\hat{\pi}) \le Ce\cdot\mathbb{E}_{\hat{\pi}}\left[\sum_h H^2\epsilon(x_h) + \epsilon(x_h,a_h) + H^2\tilde{\epsilon}({}^H\!/_2,x_h,a_h)\right]. \tag{76}$$

Summing over all $k \in [K]$,

$$\sum_{k\in[K]} v(\pi^\star) - v(\hat{\pi}_k) \le Ce\cdot\sum_{k\in[K]}\mathbb{E}_{\hat{\pi}}\left[\sum_{h\in[H]} H^2\epsilon_k(x_h) + \epsilon_k(x_h,a_h) + H^2\tilde{\epsilon}_k({}^H\!/_2,x_h,a_h)\right]. \tag{77}$$

To bound this quantity with the pigeonhole principle, we apply the Azuma-Hoeffding bound (Lemma F.4) to the martingale difference sequence defined with $Z_{k,h} := H^2\epsilon_k(x_h)+\epsilon_k(x_h,a_h)+H^2\tilde{\epsilon}_k({}^H\!/_2,x_h,a_h)$ where $|Z_{k,h} - \mathbb{E}_{\hat{\pi}_k}[Z_{k,h}]| \le 12H^2$. Therefore, under this additional event,

$$\sum_{k\in[K]} v(\pi^\star) - v(\hat{\pi}_k) \tag{78}$$

$$\le 48\cdot Ce\cdot H^2\sqrt{KH\log(2/\delta)} + Ce\cdot\sum_{k,h} H^2\epsilon_k(x_h) + \epsilon_k(x_h,a_h) + H^2\tilde{\epsilon}_k({}^H\!/_2,x_h,a_h) \tag{79}$$

$$\le 48\cdot Ce\cdot H^2\sqrt{KH\log(2/\delta)} + Ce\cdot\sum_{k,h} H^2\sqrt{\frac{C_O Y\iota}{n_k(x_h^k)\vee 1}} + \sqrt{\frac{C_T H^3\iota}{n_k(x_h^k,a_h^k)\vee 1}} + H^2\cdot\frac{2(H/2)X\iota}{n_k(x_h^k,a_h^k)\vee 1} \tag{80}$$

where $\iota = \log({}^{2X^2YAKH}\!/_\delta)$. Applying the pigeonhole principle the summations (Lemmas F.5 and F.7), we have

$$\sum_{k\in[K]} v(\pi^\star) - v(\hat{\pi}_k) \le 48\cdot Ce\sqrt{H^5K\log(2/\delta)} + 3\cdot Ce\sqrt{C_O YXH^5K\iota} + 3\cdot Ce\sqrt{C_T XAH^4K\iota} \tag{81}$$

$$+ CeH^4X^2A\iota(1+\log(K)) + Ce\left(H^2\sqrt{C_O Y\iota}HX + \sqrt{C_T H^3\iota}HXA\right) \tag{82}$$

We recall that the constants have values $C_T = 4$, $C_O = 8$, and $C = 21$. Finally we conclude that the intersection of the good events $\mathcal{E}$ and the Azuma-Hoeffding event occur simultaneously with probability at least $1-5\delta$.

C.5. Supporting results

C.5.1. Proof of Lemma C.8

Proof of Lemma C.8. The second claim follows simply by the definition of $\hat{r}$ and using the form of $\hat{\alpha}^{\hat{\pi}}_H$ and $\alpha^{\hat{\pi}}_H$ at step H. We focus on the first claim. Note that, from the recursive definitions of $\hat{\alpha}$ and α in Proposition B.1, we have

$$\hat{\alpha}^{\hat{\pi}}_{h,\tau_h}(x,a) - \alpha^{\hat{\pi}}_{h,\tau_h}(x,a) = \hat{r}(x,a) - r(x,a) + \hat{\mathcal{B}}_{\tau_h}(x,a)\left[\hat{\alpha}^{\hat{\pi}}_{h+1}\right] - \mathcal{B}_{\tau_h}(x,a)\left[\alpha^{\hat{\pi}}_{h+1}\right], \tag{83}$$

where we define the operators

$$\mathcal{B}_\tau(x,a)\left[\alpha\right] := \sum_{x',y'} O_\star(y'|x')T_\star(x'|x,a)\alpha_{\tau'}(x') \tag{84}$$

$$\hat{\mathcal{B}}_\tau(x,a)\left[\alpha\right] := \sum_{x',y'} \hat{O}(y'|x')\hat{T}(x'|x,a)\alpha_{\tau'}(x') \tag{85}$$

and we use the same convention of defining τ' as the concatenation of τ and (a,y'). Then,

$$\hat{\alpha}^{\hat{\pi}}_{h,\tau_h}(x,a) - \alpha^{\hat{\pi}}_{h,\tau_h}(x,a) = \hat{r}(x,a) - r(x,a) + \hat{\mathcal{B}}_{\tau_h}(x,a)\left[\hat{\alpha}^{\hat{\pi}}_{h+1}\right] - \mathcal{B}_{\tau_h}(x,a)\left[\alpha^{\hat{\pi}}_{h+1}\right] \tag{86}$$

$$= H\epsilon(x) + \epsilon(x,a) + \hat{\mathcal{B}}_{\tau_h}(x,a)\left[\hat{\alpha}^{\hat{\pi}}_{h+1}\right] - \mathcal{B}_{\tau_h}(x,a)\left[\alpha^{\hat{\pi}}_{h+1}\right] \tag{87}$$

$$= H\epsilon(x) + \epsilon(x,a) + \underbrace{\left(\hat{\mathcal{B}}_{\tau_h}(x,a) - \mathcal{B}_{\tau_h}(x,a)\right)\left[\hat{\alpha}^{\hat{\pi}}_{h+1}\right]}_{\textbf{(I)}} + \mathcal{B}_{\tau_h}(x,a)\left[\hat{\alpha}^{\hat{\pi}}_{h+1} - \alpha^{\hat{\pi}}_{h+1}\right] \tag{88}$$

Next, we use Lemma C.9 to get a bound on **(I)**.

Lemma C.9. *Term **(I)** is bounded above by the following quantity:*

$$\textbf{(I)} \le \frac{\mathcal{B}_{\tau_h}(x,a)\left[\hat{\alpha}^{\hat{\pi}}_{h+1} - \alpha^{\pi^\star}_{h+1}\right]}{2c} + 7H^2\tilde{\epsilon}(c,x,a) + 14H^2\tilde{\epsilon}(1,x,a) + \epsilon(x,a) + 11H^2\sum_{x'} T_\star(x'|x,a)\epsilon(x') \tag{89}$$

Hence, because $\tilde{\epsilon}(1,x,a) \le \tilde{\epsilon}(c,x,a)$ for $c \ge 1$ and we set $C = 21$, the expected difference in α-vectors is then bounded above by

$$\mathbb{E}_{\hat{\pi}}\left[\hat{\alpha}^{\hat{\pi}}_{h,\tau_h}(x_h,a_h) - \alpha^{\hat{\pi}}_{h,\tau_h}(x_h,a_h)\right] \le \mathbb{E}_{\hat{\pi}}\left[H\epsilon(x_h) + 2\epsilon(x_h,a_h) + CH^2\tilde{\epsilon}(c,x_h,a_h) + 11H^2\sum_{x'} T_\star(x'|x_h,a_h)\epsilon(x')\right] \tag{90}$$

$$+ \mathbb{E}_{\hat{\pi}}\left[\frac{\mathcal{B}_{\tau_h}(x_h,a_h)\left[\hat{\alpha}^{\hat{\pi}}_{h+1} - \alpha^{\pi^\star}_{h+1}\right]}{2c}\right] \tag{91}$$

$$+ \mathbb{E}_{\hat{\pi}}\left[\mathcal{B}_{\tau_h}(x_h,a_h)\left[\hat{\alpha}^{\hat{\pi}}_{h+1} - \alpha^{\hat{\pi}}_{h+1}\right]\right] \tag{92}$$

$$\le \mathbb{E}_{\hat{\pi}}\left[H\epsilon(x_h) + 2\epsilon(x_h,a_h) + CH^2\tilde{\epsilon}(c,x_h,a_h) + 11H^2\sum_{x'} T_\star(x'|x_h,a_h)\epsilon(x')\right] \tag{93}$$

$$+ \mathbb{E}_{\hat{\pi}}\left[\frac{\mathcal{B}_{\tau_h}(x_h,a_h)\left[\hat{\alpha}^{\hat{\pi}}_{h+1} - \alpha^{\hat{\pi}}_{h+1}\right]}{2c}\right] \tag{94}$$

$$+ \mathbb{E}_{\hat{\pi}}\left[\mathcal{B}_{\tau_h}(x_h,a_h)\left[\hat{\alpha}^{\hat{\pi}}_{h+1} - \alpha^{\hat{\pi}}_{h+1}\right]\right], \tag{95}$$

where the second line (before-and-after changes marked in red) follows because

$$\mathbb{E}_{\hat{\pi}}\left[\mathcal{B}_{\tau_h}(x_h,a_h)\left[\alpha^{\hat{\pi}}_{h+1}\right]\right] = \mathbb{E}_{\hat{\pi}}\left[V^{\hat{\pi}}_{h+1}(\tau_{h+1})\right] \tag{96}$$

$$\le \mathbb{E}_{\hat{\pi}}\left[V^{\pi^\star}_{h+1}(\tau_{h+1})\right] \tag{97}$$

$$= \mathbb{E}_{\hat{\pi}}\left[\mathcal{B}_{\tau_h}(x_h,a_h)\left[\alpha^{\pi^\star}_{h+1}\right]\right] \tag{98}$$

since $\pi^\star$ is the optimal history-dependent policy. Therefore, since $\mathbb{E}_{\hat{\pi}}\left[\mathcal{B}_{\tau_h}(x_h,a_h)\left[\alpha_{h+1}\right]\right] = \mathbb{E}_{\hat{\pi}}\left[\alpha_{h+1}(x_{h+1})\right]$, we conclude that

$$\mathbb{E}_{\hat{\pi}}\left[\hat{\alpha}^{\hat{\pi}}_{h,\tau_h}(x_h) - \alpha^{\hat{\pi}}_{h,\tau_h}(x_h)\right] = \mathbb{E}_{\hat{\pi}}\left[\hat{\alpha}^{\hat{\pi}}_{h,\tau_h}(x_h,a_h) - \alpha^{\hat{\pi}}_{h,\tau_h}(x_h,a_h)\right] \tag{99}$$

$$\le \mathbb{E}_{\hat{\pi}}\left[H\epsilon(x_h) + 2\epsilon(x_h,a_h) + CH^2\tilde{\epsilon}(c,x_h,a_h)\right] \tag{100}$$

$$+ \left(1 + \frac{1}{2c}\right)\mathbb{E}_{\hat{\pi}}\left[\hat{\alpha}^{\hat{\pi}}_{h+1,\tau_{h+1}}(x_{h+1}) - \alpha^{\hat{\pi}}_{h+1,\tau_{h+1}}(x_{h+1}) + 11H^2\epsilon(x_{h+1})\right]. \tag{101}$$

□

C.5.2. Proof of Lemma C.9

Here, we restate the bound on **(I)** before proving it.

Lemma C.9. *Term **(I)** is bounded above by the following quantity:*

$$\textbf{(I)} \leq \frac{\mathcal{B}_{\tau_h}(x,a)\left[\hat{\alpha}^{\hat{\pi}}_{h+1} - \alpha^{\pi^\star}_{h+1}\right]}{2c} + 7H^2\tilde{\epsilon}(c,x,a) + 14H^2\tilde{\epsilon}(1,x,a) + \epsilon(x,a) + 11H^2\sum_{x'} T_\star(x'|x,a)\epsilon(x') \tag{89}$$

Proof of Lemma C.9.

$$\begin{aligned}
\textbf{(I)} &= \left(\hat{\mathcal{B}}_{\tau_h}(x,a) - \mathcal{B}_{\tau_h}(x,a)\right)\left[\hat{\alpha}^{\hat{\pi}}_{h+1} - \alpha^{\pi^\star}_{h+1}\right] + \left(\hat{\mathcal{B}}_{\tau_h}(x,a) - \mathcal{B}_{\tau_h}(x,a)\right)\left[\alpha^{\pi^\star}_{h+1}\right] && (102)\\
&\leq \left(\hat{\mathcal{B}}_{\tau_h}(x,a) - \mathcal{B}_{\tau_h}(x,a)\right)\left[\hat{\alpha}^{\hat{\pi}}_{h+1} - \alpha^{\pi^\star}_{h+1}\right] + \epsilon(x,a) + (H-h)\sum_{x'}\hat{T}(x'|x,a)\epsilon(x') && (103)\\
&\leq \frac{\mathcal{B}_{\tau_h}(x,a)\left[\hat{\alpha}^{\hat{\pi}}_{h+1} - \alpha^{\pi^\star}_{h+1}\right]}{2c} + 6H^2\tilde{\epsilon}(c,x,a) + \epsilon(x,a) + 7H^2\sum_{x'}\hat{T}(x'|x,a)\epsilon(x') && (104)\\
&\leq \frac{\mathcal{B}_{\tau_h}(x,a)\left[\hat{\alpha}^{\hat{\pi}}_{h+1} - \alpha^{\pi^\star}_{h+1}\right]}{2c} + 6H^2\tilde{\epsilon}(c,x,a) + 14H^2\tilde{\epsilon}(1,x,a) + \epsilon(x,a) + 11H^2\sum_{x'} T_\star(x'|x,a)\epsilon(x'). && (105)
\end{aligned}$$

The first inequality uses Lemma C.10 and the second uses Lemma C.11. The last inequality applies Corollary C.4 using $\mathcal{E}_T^1$, which guarantees that

$$7H^2\sum_{x'}\hat{T}(x'|x,a)\epsilon(x') \leq 11H^2\sum_{x'}T_\star(x'|x,a)\epsilon(x') + 14H^2\tilde{\epsilon}(1,x,a) \tag{106}$$

since $\epsilon(x') \leq 2$ for all $x' \in \mathcal{X}$ by definition. □

C.5.3. Helpers

Lemma C.10. *If $\mathcal{E}_T$ and $\mathcal{E}_O$ hold then,*

$$\left(\hat{\mathcal{B}}_{\tau_h}(x,a) - \mathcal{B}_{\tau_h}(x,a)\right)\left[\alpha^{\pi^\star}_{h+1}\right] \leq \epsilon(x,a) + (H-h)\sum_{x'}\hat{T}(x'|x,a)\epsilon(x'). \tag{107}$$

Proof. We expand the definitions of the $\hat{\mathcal{B}}$ and $\mathcal{B}$ operators and then apply $\mathcal{E}_T$ and $\mathcal{E}_O$ directly.

$$\begin{aligned}
\left(\hat{\mathcal{B}}_{\tau_h}(x,a) - \mathcal{B}_{\tau_h}(x,a)\right)\left[\alpha^{\pi^\star}_{h+1}\right] &= \sum_{x',y'} O_\star(y'|x')\left(\hat{T}(x'|x,a) - T_\star(x'|x,a)\right)\left[\alpha^{\pi^\star}_{h+1,\tau_h'}(x')\right] && (108)\\
&\quad + \sum_{x',y'}\left(\hat{O}(y'|x') - O_\star(y'|x')\right)\hat{T}(x'|x,a)\left[\alpha^{\pi^\star}_{h+1,\tau_h'}(x')\right] && (109)\\
&\leq \epsilon(x,a) + (H-h)\sum_{x'}\hat{T}(x'|x,a)\epsilon(x'). && (110)
\end{aligned}$$

□

Lemma C.11. *If $\mathcal{E}_T^c$ and $\mathcal{E}_O$ hold, then*

$$\begin{aligned}
\left(\hat{\mathcal{B}}_{\tau_h}(x,a) - \mathcal{B}_{\tau_h}(x,a)\right)\left[\hat{\alpha}^{\hat{\pi}}_{h+1} - \alpha^{\pi^\star}_{h+1}\right] &\leq \frac{\mathcal{B}_{\tau_h}(x,a)\left[\hat{\alpha}^{\hat{\pi}}_{h+1} - \alpha^{\pi^\star}_{h+1}\right]}{2c} + 6H^2\tilde{\epsilon}(c,x,a) && (111)\\
&\quad + 6H^2\sum_{x'}\hat{T}(x'|x,a)\epsilon(x'). && (112)
\end{aligned}$$

Proof. For convenience, let $\alpha := \hat{\alpha}^{\hat{\pi}}$ and $\alpha' := \alpha^{\pi^\star}$. Observe that Proposition B.1 and Lemma C.7 guarantee that $|\alpha_{h+1,\tau'_h} - \alpha'_{h+1,\tau'_h}| \le 6H(H-h)$. Then,

$$\left(\hat{\mathcal{B}}_{\tau_h}(x,a) - \mathcal{B}_{\tau_h}(x,a)\right)\left[\alpha_{h+1} - \alpha'_{h+1}\right] \tag{113}$$

$$= \sum_{x'}\left(\hat{T}(x'|x,a) - T_\star(x'|x,a)\right)\sum_{y'} O_\star(y'|x')\left(\alpha_{h+1,\tau'_h}(x') - \alpha'_{h+1,\tau'_h}(x')\right) \tag{114}$$

$$+ \sum_{x',y'} \hat{T}(x'|x,a)\left(\hat{O}(y'|x') - O_\star(y'|x')\right)\left(\alpha_{h+1,\tau'_h}(x') - \alpha'_{h+1,\tau'_h}(x')\right) \tag{115}$$

$$\le \sum_{x'}\left(\hat{T}(x'|x,a) - T_\star(x'|x,a)\right)\sum_{y'} O_\star(y'|x')\left(\alpha_{h+1,\tau'_h}(x') - \alpha'_{h+1,\tau'_h}(x')\right) \tag{116}$$

$$+ 6H(H-h)\sum_{x',y'} \hat{T}(x'|x,a)\epsilon(x') \tag{117}$$

$$\le \frac{\sum_{x'} T_\star(x'|x,a)\sum_{y'} O_\star(y'|x')\left(\alpha_{h+1,\tau'_h}(x') - \alpha'_{h+1,\tau'_h}(x')\right)}{2c} \tag{118}$$

$$+6H(H-h)\min\left\{2, \frac{2cX\log(X^2AKH/\delta)}{n(x,a)}\right\} \tag{119}$$

$$+ 6H(H-h)\sum_{x'} \hat{T}(x'|x,a)\epsilon(x') \tag{120}$$

$$\le \frac{\mathcal{B}_{\tau_h}(x,a)\left[\alpha_{h+1} - \alpha'_{h+1}\right]}{2c} + 6H^2\tilde{\epsilon}(c,x,a) + 6H^2\sum_{x'}\hat{T}(x'|x,a)\epsilon(x'). \tag{121}$$

The first inequality uses $\mathcal{E}_O$ to bound the total variation distance between $O_\star$ and $\hat{O}$ (before-and-after changes marked in red). The second inequality uses $\mathcal{E}_T^c$ along with Corollary C.4 by setting $g(x') = \sum_{y'} O(y'|x')\left(\alpha_{h+1,\tau'_h}(x') - \alpha'_{h+1,\tau'_h}(x')\right)$ (before-and-after changes marked in blue). The last inequality simply uses the definition of $\mathcal{B}$ and $\tilde{\epsilon}(c,x,a)$ and the fact that $H(H-h) \le H^2$.

□

C.6. First steps towards function approximation

A natural follow-up question is whether HOP-B can be easily generalized to incorporate function approximation. While we leave in depth discussion of a much more general form of function approximation to Section 6, we remark that it is easy to replace the tabular estimation of $O_\star$ with function approximation as long as the latent states are tabular.

Consider a function class $\Theta \subseteq (\mathcal{X} \to \Delta(\mathcal{Y}))$. For simplicity assume that Θ is finite, in which case we expect the complexity of Θ to be measured as the log-cardinality $\log(|\Theta|)$, as is standard. Assuming that Θ realizes $O_\star$ ($O_\star \in \Theta$) and it is proper ($O(\cdot|x) \in \Delta(\mathcal{Y})$ for all $x \in \mathcal{X}$ and $O \in \Theta$), one can update $\hat{O}_k$ via maximum likelihood estimation (MLE):

$$\hat{O}_{k+1} = \arg\max_{O\in\Theta} \sum_{\ell\in[k],h\in[H]} \log O(y_h^\ell|x_h^\ell).$$

Then, we change the bonuses to be

$$\epsilon_k(x,a) = \min\left\{2, \sqrt{\frac{C'_T X\log\left(X^2AKH\right)}{n_k(x,a)}}\right\} \tag{122}$$

$$\epsilon_k(x) = \min\left\{2, \sqrt{\frac{C'_O\log(|\Theta|XK/\delta)}{n_k(x)}}\right\}, \tag{123}$$

where $C'_T = 8$ and $C'_O = 8$ and the optimistic reward function is changed to

$$\hat{r}_k(x,a) = r(x,a) + 3H\epsilon_k(x,a) + H\epsilon_k(x). \tag{124}$$

Note that the algorithm still requires only point estimates of $O_\star$ as opposed to maintaining a version space. These changes yield the following bound:

Proposition C.12. *Let $\mathcal{M}$ be a HOMDP model with X latent states. With probability at least $1-3\delta$, HOP-B with emission function class Θ outputs a sequence of policies $\hat{\pi}_1, \dots, \hat{\pi}_K$ such that*

$$\mathrm{Reg}(K) = \mathcal{O}\left(\sqrt{H^5XK\left(\log(|\Theta|)+\iota\right)} + \sqrt{H^5X^2AK\iota}\right) \tag{125}$$

where $\iota = \log(2X^2AKH/\delta)$.

We see that the original Y dependence is replaced with the complexity $\log(|\Theta|)$. An interesting observation of this result is that we do not have to tailor the exploration to the type of function approximator used for $O_\star$ aside from adjustment of the bonus magnitude. The class Θ can also be completely arbitrary and need not satisfy any further structural conditions besides realizability and learnability for the MLE (i.e. manageable complexity).

We finally remark that the dependence on X is worse than the purely tabular case. This is due to an alternative technical approach (akin to the difference between the UCRL bound of Auer et al. (2008) and the improved version of Azar et al. (2017)). It is possible to apply our original technique to this case as well, but, in contrast to MDPs, this would yield a $\log(Y)$ factor due to a union bound over histories, which is not ideal. We believe the simplicity of this analysis and ability to handle infinite Y is a more desirable choice.

C.6.1. Proof of Proposition C.12

We define new high probability events for the count-based estimate $\hat{T}_k$ and maximum likelihood estimate $\hat{O}_k$:

$$\mathcal{E}_T = \left\{\forall k \in [K], x \in \mathcal{X}, a \in \mathcal{A},\ \|T(\cdot|x,a) - \hat{T}_k(\cdot|x,a)\|_1 \le \sqrt{\frac{C'_T X \log(X^2AKH/\delta)}{n_k(x,a)}}\right\} \tag{126}$$

$$\mathcal{E}_O = \left\{\forall k \in [K], x \in \mathcal{X},\ \|O(\cdot|x) - \hat{O}_k(\cdot|x)\|_1 \le \sqrt{\frac{C'_O \log(XK|\Theta|/\delta)}{n_k(x)}}\right\}. \tag{127}$$

with $C'_T = 8$ and $C'_O = 8$. $P(\mathcal{E}_T) \ge 1-\delta$ follows the same proof as Lemma C.5 and $P(\mathcal{E}_O) \ge 1-\delta$ follows the same proof as Lemma E.2 but applied to the emission function (see also Theorem 21 of Agarwal et al. (2020)). This guarantees that, with probability at least $1-\delta$,

$$n_k(x)\|O(\cdot|x_h^\ell) - \hat{O}_k(\cdot|x_h^\ell)\|_1^2 \le \sum_{\ell\in[k-1],h\in[H]} \|O(\cdot|x_h^\ell) - \hat{O}_k(\cdot|x_h^\ell)\|_1^2 \tag{128}$$

$$\le 8\log(K|\Theta|/\delta) \tag{129}$$

for all $k \in [K]$. Rearranging ensures the claim on $P(\mathcal{E}_O)$. Assuming these events hold, we show that this ensures optimism of the α-vectors as before.

Let $\hat{\alpha}$ denote the α-vector for the estimated model $\hat{T}$ and $\hat{O}$ and let α be the one for the true model.

Lemma C.13. *Let the above events hold. Then, $\hat{\alpha}^\pi_{k,h,\tau_h} \ge \alpha^\pi_{h,\tau_h}(x) + H\epsilon_k(x)$*

Proof. For now, we omit the subscript notation denoting the round k. By the above events, we have that

$$\hat{\alpha}^\pi_{H,\tau_H}(x) = \sum_a \pi(a|\tau_h)\left(r(x,a) + H\epsilon(x) + 3H\epsilon(x,a)\right), \tag{130}$$

which means that

$$\hat{\alpha}^\pi_{H,\tau_H}(x) - \alpha^\pi_{H,\tau_H}(x) \ge H\epsilon(x) + 3H\epsilon(x,a) \tag{131}$$

$$\ge H\epsilon(x) \tag{132}$$

Inductively, assume that $\hat{\alpha}^{\pi}_{h+1,\tau_{h+1}}(x) \geq \alpha^{\pi}_{h+1,\tau_{h+1}}(x) + H\epsilon(x)$. Then, using recursive definitions of the α-vectors,

$$\begin{aligned}
\hat{\alpha}^{\pi}_{h,\tau_h}(x,a) - \alpha^{\pi}_{h,\tau_h}(x,a) &= \hat{r}(x,a) + \sum_{x',y'} \hat{T}(x'|x,a)\hat{O}(y'|x')\hat{\alpha}^{\pi}_{h+1,\tau'_h}(x') && (133)\\
&\quad - r(x,a) - \sum_{x',y'} T_\star(x'|x,a)O_\star(y'|x')\alpha^{\pi}_{h+1,\tau'_h}(x') && (134)\\
&\geq \hat{r}(x,a) - r(x,a) && (135)\\
&\quad + \sum_{x',y'} \hat{T}(x'|x,a)\hat{O}(y'|x') \left(\alpha^{\pi}_{h+1,\tau'_h}(x') + H\epsilon(x')\right) && (136)\\
&\quad - \sum_{x',y'} T_\star(x'|x,a)O_\star(y'|x')\alpha^{\pi}_{h+1,\tau'_h}(x') && (137)\\
&= \hat{r}(x,a) - r(x,a) && (138)\\
&\quad + \sum_{x',y'} \left(\hat{T}(x'|x,a) - T_\star(x'|x,a)\right) \hat{O}(y'|x')\alpha^{\pi}_{h+1,\tau'_h}(x') && (139)\\
&\quad + \sum_{x',y'} T_\star(x'|x,a)(x')\alpha^{\pi}_{h+1,\tau'_h} \left(\hat{O}(y'|x') - O_\star(y'|x')\right) && (140)\\
&\quad + H \sum_{x',y'} \hat{T}(x'|x,a)\hat{O}(y'|x')\epsilon(x'), && (141)
\end{aligned}$$

where τ'_h is again the concatenation of τ_h with y' and a. Now, we can lower bound the above using the total variation distance:

$$\begin{aligned}
\hat{\alpha}^{\pi}_{h,\tau_h}(x,a) - \alpha^{\pi}_{h,\tau_h}(x,a) &\geq \hat{r}(x,a) - r(x,a) && (142)\\
&\quad - H\|\hat{T}(\cdot|x,a) - T_\star(\cdot|x,a)\|_1 && (143)\\
&\quad - H \sum_{x'} T_\star(x'|x,a)\|\hat{O}(\cdot|x') - O_\star(\cdot|x')\|_1 && (144)\\
&\quad + H \sum_{x'} \hat{T}(x'|x,a)\epsilon(x'). && (145)\\
& && (146)
\end{aligned}$$

Recognizing that $\epsilon(x') \leq 2$ by definition, the last term is lower bounded as

$$H \sum_{x'} \hat{T}(x'|x,a)\epsilon(x') \geq -2H\|\hat{T}(\cdot|x,a) - T_\star(\cdot|x,a)\|_1 + H \sum_{x'} T_\star(x'|x,a)\epsilon(x') \tag{147}$$

Putting these all together, we have

$$\begin{aligned}
\hat{\alpha}^{\pi}_{h,\tau_h}(x,a) - \alpha^{\pi}_{h,\tau_h}(x,a) &\geq \hat{r}(x,a) - r(x,a) && (148)\\
&\quad - 3H\|\hat{T}(\cdot|x,a) - T_\star(\cdot|x,a)\|_1 && (149)\\
&\quad - H \sum_{x'} T_\star(x'|x,a) \left(\|\hat{O}(\cdot|x') - O_\star(\cdot|x')\|_1 - \epsilon(x')\right) && (150)\\
&\geq \hat{r}(x,a) - r(x,a) && (151)\\
&\quad - 3H\|\hat{T}(\cdot|x,a) - T_\star(\cdot|x,a)\|_1 && (152)\\
&\geq H\epsilon(x) + 3H\epsilon(x,a) - 3H\|\hat{T}(\cdot|x,a) - T_\star(\cdot|x,a)\|_1 && (153)\\
&\geq H\epsilon(x) && (154)
\end{aligned}$$

Applying this inductive argument backwards along $h = H, \ldots, 1$ gives the result.

□

Lemma C.14. *Let the above events hold. Then, for any history-dependent policy π, it holds that $\hat{v}_k(\pi) - v(\pi) \geq 0$, where $\hat{v}_k$ is the policy value under the model $\hat{T}_k$, $\hat{O}_k$, and $\hat{r}_k$.*

Proof. The proof is immediate from the α-vector representation of the value functions and Lemma C.13:

$$\begin{aligned}
\hat{v}_k(\pi) - v(\pi) &= \sum_{x_1, y_1} \hat{O}(y_1|x_1)\rho(x_1)\hat{\alpha}^{\pi}_{k,1,\tau_1}(x_1) - \sum_{x_1, y_1} O_\star(y_1|x_1)\rho(x_1)\alpha^{\pi}_{1,\tau_1}(x_1) && (155)\\
&= \sum_{x_1, y_1} \left(\hat{O}_k(y_1|x_1) - O_\star(y_1|x_1)\right)\rho(x_1)\alpha^{\pi}_{1,\tau_1}(x_1) && (156)\\
&\quad + \sum_{x_1, y_1} \hat{O}(y_1|x_1)\rho(x_1)\left(\hat{\alpha}^{\pi}_{k,1,\tau_1}(x_1) - \alpha^{\pi}_{1,\tau_1}(x_1)\right) && (157)\\
&\geq -\sum_{x_1} H\rho(x_1)\epsilon_k(x_1) + \sum_{x_1, y_1} H\hat{O}(y_1|x_1)\rho(x_1)\epsilon_k(x_1) && (158)\\
&= 0. && (159)
\end{aligned}$$

□

We are now ready to prove the result. Let $\hat{\mathbb{E}}_k$ and $\hat{P}_k$ denote the expectation and measure under the learned model at round k. Then,

$$\begin{aligned}
\mathrm{Reg}(K) &= \sum_{k\in[K]} v(\pi^\star) - v(\hat{\pi}_k) && (160)\\
&\leq \sum_{k\in[K]} \hat{v}_k(\hat{\pi}_k) - v(\hat{\pi}_k) && (161)\\
&= \sum_{k\in[K]} \hat{\mathbb{E}}_{k,\hat{\pi}_k}\left[\sum_{h\in[H]} \hat{r}_k(x_h, a_h)\right] - \mathbb{E}_{\hat{\pi}_k}\left[\sum_{h\in[H]} r(x_h, a_h)\right] && (162)\\
&= \sum_{k\in[K]} \hat{\mathbb{E}}_{k,\hat{\pi}_k}\left[\sum_{h\in[H]} r(x_h, a_h) + H\epsilon_k(x_h) + 3H\epsilon_k(x_h, a_h)\right] - \mathbb{E}_{\hat{\pi}_k}\left[\sum_{h\in[H]} r(x_h, a_h)\right] && (163)\\
&\leq \sum_{k\in[K]} H\|\hat{P}_{k,\hat{\pi}_k} - P_{\hat{\pi}_k}\|_1 + 3H\hat{\mathbb{E}}_{k,\hat{\pi}_k}\left[\sum_{h\in[H]} \epsilon_k(x_h) + \epsilon_k(x_h, a_h)\right]. && (164)
\end{aligned}$$

Since $\epsilon_k(x)$ and $\epsilon_k(x, a)$ are no greater than 2, we can change distributions in the last term of the previous display to get

$$\mathrm{Reg}(K) \leq \sum_{k\in[K]} 13H^2\|\hat{P}_{k,\hat{\pi}_k} - P_{\hat{\pi}_k}\|_1 + 3H\mathbb{E}_{\hat{\pi}_k}\left[\sum_{h\in[H]} \epsilon_k(x_h) + \epsilon_k(x_h, a_h)\right]. \quad (165)$$

To bound the remaining terms, we rely on the Azuma-Hoeffding inequality (Lemma F.4) with probability at least $1 - \delta$ and the simulation lemma (Lemma E.1).

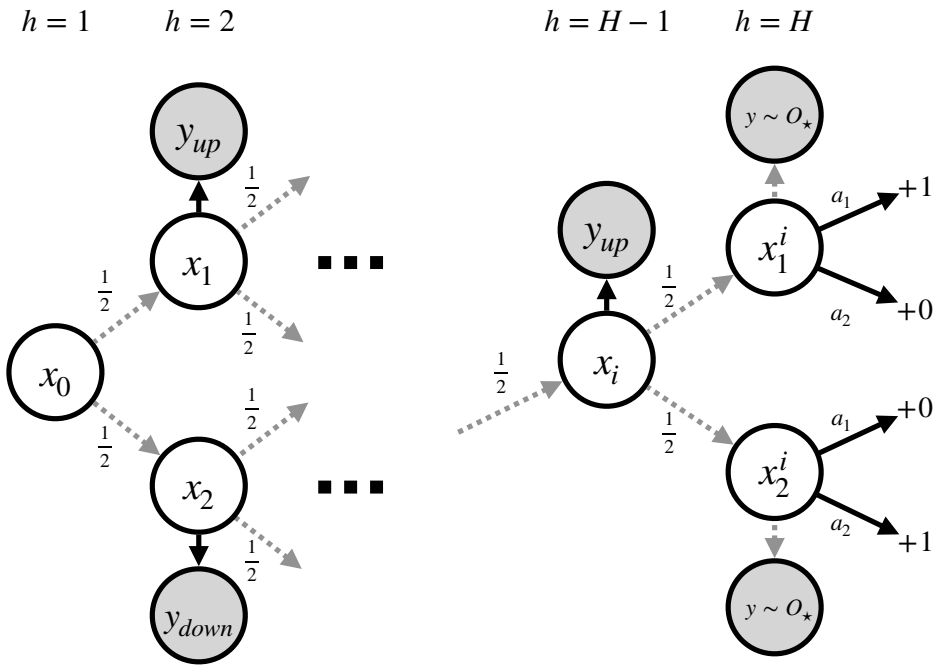

Figure 3: Hard instance of POMDP. The latent state space is a binary tree starting at x_0 and the learner traverses a layer each time step. The root and and second layer depicted on the left. Observations about which direction it traversed are revealed at each step. In the last and second to last layer, one of many triplets like Figure 2 is encountered (only one depicted in this figure on the right). The policy has no control until the last layer.

The regret can then further be bounded as

$$\mathrm{Reg}(K) \le \sum_{k\in[K]} 13H^2\mathbb{E}_{\hat{\pi}_k}\left[\sum_{h\in[H]} \epsilon_k(x_h)+\epsilon_k(x_h,a_h)\right] + 3H\mathbb{E}_{\hat{\pi}_k}\left[\sum_{h\in[H]} \epsilon_k(x_h)+\epsilon_k(x_h,a_h)\right] \tag{166}$$

$$= 16H^2 \sum_{k\in[K],h\in[H]} \mathbb{E}_{\hat{\pi}_k}\left[\epsilon_k(x_h)+\epsilon_k(x_h,a_h)\right] \tag{167}$$

$$\le 64H^2\sqrt{KH\log(2/\delta)} + 16H^2\sum_{k,h}\epsilon_k(x_h^k)+\epsilon_k(x_h^k,a_h^k) \tag{168}$$

$$\le 64H^2\sqrt{KH\log(2/\delta)} + 16H^2\sum_{k,h}\sqrt{\frac{C_O'(\log(|\Theta|)+\iota)}{\max\{1,n_k(x_h^k)\}}} + \sqrt{\frac{C_T'X\iota}{\max\{1,n_k(x_h^k,a_h^k)\}}}, \tag{169}$$

To bound this final term, as before, we appeal to pigeonhole principle (Lemma F.5):

$$\mathrm{Reg}(K) \lesssim \sqrt{H^5K\log(2/\delta)} + \sqrt{H^5XK\left(\log(|\Theta|)+\iota\right)} + \sqrt{H^5X^2AK\iota} \tag{170}$$

$$+ H^3X\sqrt{\log(|\Theta|)+\iota} + H^3XA\sqrt{X\iota}. \tag{171}$$

By a union bound on the events $\mathcal{E}_T$ and $\mathcal{E}_O$ and the Azuma-Hoeffding event, we conclude that this occurs with probability at least $1-3\delta$.

D. Proof of Lower Bound Theorem 5.1

In this section, we formally prove the information-theoretic lower bound of Theorem 5.1. We take the standard minimax approach: design a class of problem instances $\mathcal{U}$ such that, for any algorithm that generates a policy $\hat{\pi}$ over K episodes of interaction, there is always a problem instance $u \in \mathcal{U}$ with

$$\mathbb{E}_u\left[v(\pi^\star) - v(\hat{\pi})\right] \ge \epsilon$$

for some ϵ that we will try to control. The expectation is over the algorithm and data generated under the instance u.

While the intuition for the lower bound in the main paper with $X = 3$ is useful for understanding how the Y dependence appears in the lower bound, it is not immediate to generalize. Note that it is not sufficient to take the naive approach of simply maintaining $H = 1$ and increasing the number of latent states. In such cases, the possible problem instances will either have very low loss separation or be very easy to test and differentiate. This would mean an algorithm could easily either find out how to act optimally or not acting optimally is "close enough." The issue is that the posterior distribution over $\mathcal{X}$ given an observation would end up being very close to uniform (to avoid being able to simply test to distinguish between instances). However, this allows for a potentially large margin of error for any policy since even the optimal policy will struggle greatly to achieve high rewards regardless.

This issue can be resolved by leveraging the history to reduce the problem to solving many (on the order X) 3-latent state subproblems. Consider the sketch POMDP in Figure 3, which is the full version of the on the one in Figure 2. Figure 3 depicts a binary tree in which the policy randomly traverses the nodes to the leaves of the last layer H. The last and second to last layers collectively make up a swath of 3-latent state problems. The catch is that, as the policy traverses the nodes, the observations reveal its exact state in the tree up to layer $H - 1$ by revealing whether it traversed to the upper or lower child at each step with $\{y_{\text{up}}, y_{\text{down}}\}$. Thus it can exactly decode its state at layer $H - 1$ and find out which of the 3-latent state problems it is in. At layer $H - 1$, it is faced with one of the possible 3-latent state problems and it must act optimally.

We will design the emission function such that it is difficult to decode whether the learner is in the upper or lower child in the same way as we described for the case when $X = 3$. Thus, the learner will have to visit a given $h = H - 1$ layer parent at least $\Omega(Y)$ times before learning the optimal policy for that parent. Furthermore, it must learn the optimal policy for at least a constant fraction of the parents (size $\Omega(X)$) to compete with the full optimal policy. Together this yields total interactions on the order of $\Omega(XY)$.

Theorem 5.1. *Fix $\epsilon \leq 1/64$ and $X, Y \in \mathbb{N}$ such that $Y \geq 6$, $(X + 1) \geq 128 \log 2$. For any algorithm $\mathfrak{A}$ producing a policy $\hat{\pi}$ in K episodes of interaction, there exists a HOMDP with the aforementioned cardinalities and $H \asymp \log_2(X)$ and $A = 2$ such that $\mathfrak{A}$ needs*

$$K = \Omega\left(XY/\epsilon^2\right)$$

to guarantee $\mathbb{E}\left[v(\pi^\star) - v(\hat{\pi})\right] \leq \epsilon$, where the expectation is taken over randomness in the data and algorithm.

D.1. Construction of instance class

Note that the preconditions ensure that $Y(X + 1) \geq 512 \log 2$.

The learner starts deterministically at x_0. Without loss of generality, we assume that $X + 1$ is a power of 2 and Y is even. Otherwise, we can reduce $X + 1$ to the nearest power of 2 and reduce Y to the nearest even number by at most losing a constant factor in the bound. There are a total of $H = \log_2(X + 1)$ timesteps. Before the Hth timestep, rewards are all zero and the learner transitions uniformly randomly from the current latent state to either the upper or lower child latent state in the next layer, regardless of the action (see Figure 3). An observation y_{up} or y_{down} is revealed indicating whether the learner is currently in the upper or lower child of the parent latent state. Hence, for $h \leq H - 1$, the latent state can be exactly decoded given the history of observations. We assume that all states in the final layer $h = H$ then transition to a dummy state such as x_0 at $H + 1$.

We denote the states of the final layer $h = H$ by $\mathcal{X}'$. Note that $\mathcal{X}'$ consists of $X' := \frac{X+1}{2}$ latent states. States in $\mathcal{X}'$ can be grouped into $\frac{X'}{2}$ groups of 2 where the states in a group share the same parent. Let x_i^1 and x_i^2 be the upper and lower children of the same parent state x_i, respectively. The reward function is defined as

$$r(x_i^1, a) = \begin{cases} 1 & a = a_1 \\ 0 & a = a_2 \end{cases} \tag{172}$$

$$r(x_i^2, a) = \begin{cases} 0 & a = a_1 \\ 1 & a = a_2 \end{cases} \tag{173}$$

This is duplicated for all the parent-child triplets (x_i, x_i^1, x_i^2) in the second to last and last layers. The emission function in the final layer is different depending on the parent x_i and the child. For a child latent state $x \in \mathcal{X}'$, we design $O_\star(\cdot|x)$ to be

supported on $\mathcal{Y}' := \mathcal{Y} \setminus \{y_{\text{up}}, y_{\text{down}}\}$ with cardinality $Y' = |\mathcal{Y}'|$, which will be specified presently. The instances will vary based on the selection of $O_\star$.

D.2. Selection of emission function

We construct the instances by perturbing the probabilities in $O_\star$ so that they deviate slightly from uniform. To ensure that probabilities properly sum to 1, we will split $\mathcal{Y}'$ into equal partitions $\mathcal{Y}'_+$ and $\mathcal{Y}'_-$ each of size $\frac{|Y'|}{2}$. We can construct a bijection such that for any $y \in \mathcal{Y}'_-$ there is a "mirror" y_+ in $\mathcal{Y}'_+$. An instance in the class will be specified by some vector $u \in \{-1, 1\}^{\frac{X'Y'}{4}}$ which determines the observation matrix. We will denote the observation matrix for instance u with $O_{\star,u}$. Fix $\epsilon > 0$. We index into the vector u with an observation in $y \in \mathcal{Y}'_+$ and a parent latent state $x_i \in \mathcal{X}$ in the second to last layer. Let x_i^1 and x_i^2 be leaf latent states that share the same parent x_i. Note that this is valid since the child states (as we construct them) are distinct for each parent. Then for x_i^1, define

$$O_{\star,u}(y|x_i^1) = \frac{1 + u(y, x_i)\epsilon}{Y'} \quad \forall y \in \mathcal{Y}'_+, \tag{174}$$

$$O_{\star,u}(y|x_i^1) = \frac{1 - u(y_+, x_i)\epsilon}{Y'} \quad \forall y \in \mathcal{Y}'_-, \tag{175}$$

and for x_2,

$$O_{\star,u}(y|x_i^2) = \frac{1 - u(y, x_i)\epsilon}{Y'} \quad \forall y \in \mathcal{Y}'_+, \tag{176}$$

$$O_{\star,u}(y|x_i^2) = \frac{1 + u(y_+, x_i)\epsilon}{Y'} \quad \forall y \in \mathcal{Y}'_-. \tag{177}$$

We will also use P_u and $\mathbb{E}_u$ to denote the measure and expectation, respectively, under instance u. We can verify that, conditioned on the the parent x_i, the distribution over $\mathcal{Y}'$ is uniform:

$$P_u(y|x_i) = P(x_i^1|x_i)O_{\star,u}(y|x_i^1) + P(x_i^2|x_i)O_{\star,u}(y|x_i^2) \tag{178}$$

$$= \frac{O_{\star,u}(y|x_i^1) + O_{\star,u}(y|x_i^2)}{2} = \frac{1}{Y'}. \tag{179}$$

It is also worth noting that, conditioned on the parent x_i (equivalently, on the history $y_{1:H-1}$), the posterior is:

$$P_u(x_i^1|y_H, x_i) = \frac{1 + u(y_H, x_i)\epsilon}{2} \quad \forall y \in \mathcal{Y}'_+ \tag{180}$$

$$P_u(x_i^1|y_H, x_i) = \frac{1 - u((y_H)_+, x_i)\epsilon}{2} \quad \forall y \in \mathcal{Y}'_-. \tag{181}$$

Furthermore, by Lemma 4.7 of Massart (2007), there exists $\mathcal{U} \in \{-1, 1\}^{X'Y'/4}$ such that $|\mathcal{U}| \geq \exp(Y'X'/32)$ and $\|u - u'\|_1 \geq \frac{X'Y'}{8}$ for all $u, u' \in \mathcal{U}$ such that $u \neq u'$.

D.3. Separability condition

We now show in this construction that no history-dependent policy can perform well on all instances in $\mathcal{U}$ simultaneously. Let Π be the class of all history-dependent, deterministic policies. Let $u \in \mathcal{U}$ be fixed. It is clear that the optimal policy for instance u chooses action $\hat{a}$ such that $r(\hat{x}, \hat{a}) = 1$ where $\hat{x} = \arg\max_x P_u(x|y_{1:H}) \in \mathcal{X}'$ maximizes the posterior. We denote this policy by $\pi_u^\star$.

Consider an arbitrary π. Recall that, by construction, there is a bijection between $y_{1:H-1}$ and the parent latent states at layer $H-1$. To avoid notational clutter, we will now denote this by z (which we have previously written as x_i). Thus, we can equivalently write π as a function of z and the last observation y_H. Define

$$v_u(\pi|z) = \mathbb{E}_u\left[r(x, \pi(y_H, z)) \mid z\right] \tag{182}$$

as the conditional value of π given it has reached the parent z in instance u. A straightforward calculation shows that

$$v_u(\pi_u^\star|z) - v_u(\pi|z) = \epsilon \cdot \frac{N_z(\pi, \pi_u^\star)}{Y'} \tag{183}$$

where $N_z(\pi, \pi_u^\star) := \sum_{y \in \mathcal{Y}'} \mathbf{1}\{\pi_u^\star(y,z) \neq \pi(y,z)\}$ is the number of observations on which π and $\pi_u^\star$ disagree given parent z. The sub-optimality gap on the full instance is simply the average over the $\frac{X'}{2}$ parents:

$$v_u(\pi_u^\star) - v_u(\pi) = \epsilon \sum_z \frac{2N_z(\pi, \pi_u^\star)}{Y'X'} = \frac{2\epsilon N(\pi, \pi_u^*)}{Y'X'}, \tag{184}$$

where $N(\pi, \pi_u^*) = \sum_z \sum_{y \in \mathcal{Y}'} \mathbf{1}\{\pi_u^\star(y,z) \neq \pi(y,z)\}$ is the total number of disagreements at layer H. Then, for any $u, u' \in \mathcal{U}$ such that $u \neq u'$,

$$\underbrace{v_u(\pi_u^\star) - v_u(\pi)}_{\text{error on instance } u} + \underbrace{v_{u'}(\pi_{u'}^\star) - v_{u'}(\pi)}_{\text{error on instance } u'} = \frac{2\epsilon}{Y'X'}\left(N(\pi, \pi_u^\star) + N(\pi, \pi_{u'}^\star)\right) \tag{185}$$

$$\geq \frac{2\epsilon N(\pi_u^\star, \pi_{u'}^\star)}{Y'X'}. \tag{186}$$

Finally, we recall that $\mathcal{U}$ is such that $\|u - u'\|_1 \geq \frac{X'Y'}{8}$, which ensures that u and u' differ on at least $\frac{X'Y'}{16}$ elements, which implies that $N(\pi_u^\star, \pi_{u'}^\star) \geq \frac{X'Y'}{16}$. Thus, we have

$$v_u(\pi_u^\star) - v_u(\pi) + v_{u'}(\pi_{u'}^\star) - v_{u'}(\pi) \geq \frac{\epsilon}{8} \tag{187}$$

D.4. Fano's inequality application

Thus far, we have detailed the instance class and shown that no policy can achieve error less than $\frac{\epsilon}{16}$ on more than one instance. To complete the proof, we apply Fano's inequality to show that these instances are essentially indistinguishable.

Let $\hat{\pi}$ be the output of any algorithm $\mathfrak{A}$ that samples from a POMDP instance over K episodes (which a random variable dependent on the instance in which it is run). We have

$$\max_{u \in \mathcal{U}} \mathbb{E}_u\left[v_u(\pi_u^\star) - v_u(\hat{\pi})\right] \geq \frac{\epsilon}{16} \inf_{\Psi} \frac{1}{|\mathcal{U}|} \sum_{u \in \mathcal{U}} P_u\left(\Psi \neq u\right) \tag{188}$$

where Ψ is a data-dependent test function, the $\inf$ is taken over all measurable tests, P_u denotes the measure under instance u. By Fano's inequality,

$$\max_{u \in \mathcal{U}} \mathbb{E}_u\left[v_u(\pi_u^\star) - v_u(\hat{\pi})\right] \geq \frac{\epsilon}{16}\left(1 - \frac{\max_{u \neq u'} D_{KL}\left(P_u \| P_{u'}\right) + \log 2}{\log |\mathcal{U}|}\right) \tag{189}$$

where P_u and $P_{u'}$ are measures under instances u and u', respectively. Note that these are dependent on the algorithm $\mathfrak{A}$, which determines which actions to take over the K episodes. That is, the probability of taking action a_h^k in round k at step h is $\mathfrak{A}(a_h^k | \bar{\tau}^{1:k-1}, \bar{\tau}_h^k)$ where we recall that $\bar{\tau}_h^k = (x_{1:h}, y_{1:h}, a_{1:h-1})$ is the partial trajectory and $\bar{\tau}^k$ is the full trajectory, both containing the latent states. Crucially, note that we allow $\mathfrak{A}$ to be dependent on the latent states. Note that, for HOMDP model, we usually assume that this is further reduced to only dependence on historical observations and actions within the current partial trajectory so that $\bar{\tau}_h^k$ becomes τ_h^k in the conditional part. However, this generality allows us to capture algorithms that also can access the underlying state even during training deployments.

The chain rule of the KL divergence gives us the following decomposition in terms of conditional KL divergences:

$$D_{KL}(P_u \| P_{u'}) = \sum_{k \in [K]} \mathbb{E}_{\bar{\tau}^{1:k-1}}\left[D_{KL}\left(P_u(\bar{\tau}^k) \| P_{u'}(\bar{\tau}^k) \mid \bar{\tau}^{1:k-1}\right)\right].$$

where we abuse notation slightly and let $P_u(\bar{\tau}^k)$ denote the distribution over trajectory $\bar{\tau}^k$. Then, the individual terms are also written as

$$\mathbb{E}_{\bar{\tau}^{1:k-1}}\left[D_{KL}\left(P_u(\bar{\tau}^k) \| P_{u'}(\bar{\tau}^k) \mid \bar{\tau}^{1:k-1}\right)\right] = \mathbb{E}_{\bar{\tau}^{1:k-1}}\left[\sum_{\bar{\tau}^k} P_u(\bar{\tau}^k \mid \bar{\tau}^{1:k-1}) \log \frac{P_u(\bar{\tau}^k \mid \bar{\tau}^{1:k-1})}{P_{u'}(\bar{\tau}^k \mid \bar{\tau}^{1:k-1})}\right]. \tag{190}$$

Observe that the conditional probability of a trajectory is

$$P_u(\bar{\tau}^k \mid \bar{\tau}^{1:k-1}) = P_u(x_1^k) \prod_{h=1}^{H} O_{\star,u}(y_h^k|x_h^k)\mathfrak{A}(a_h|\bar{\tau}_h^k, \bar{\tau}^{1:k-1})T_{\star,u}(x_{h+1}^k|x_h^k, a_h^k) \tag{191}$$

Between the instances u and u', everything is the same (including $\mathfrak{A}$) except for $O_{\star,u}(\cdot|x_H^k)$ and $O_{\star,u'}(\cdot|x_H^k)$ in the last layer by construction. Furthermore, we have that $P(x_H^k|\bar{\tau}^{1:k-1}) = P(x_H^k) = \frac{1}{X'}$ since the policy has no control over the first $H-1$ steps. Therefore, the conditional KL divergence for any k becomes:

$$\begin{aligned}
\mathbb{E}_{\bar{\tau}^{1:k-1}}\left[D_{KL}\left(P_u(\bar{\tau}^k)\|P_{u'}(\bar{\tau}^k) \mid \bar{\tau}^{1:k-1}\right)\right] &= \sum_{y_H, x_H} P_u(x_H|\bar{\tau}_H^{1:k-1}) \log \frac{O_{\star,u}(y_H|x_H)}{O_{\star,u'}(y_H|x_H)} && (192)\\
&= \frac{1}{X'} \sum_{x_H \in \mathcal{X}'} \sum_{y_H} \log \frac{O_{\star,u}(y_H|x_H)}{O_{\star,u'}(y_H|x_H)} && (193)\\
&= \frac{1}{X'} \sum_{x \in \mathcal{X}'} \sum_{y \in \mathcal{Y}'_-} O_{\star,u}(y|x) \log \frac{O_{\star,u}(y|x)}{O_{\star,u'}(y|x)} + O_{\star,u}(y_+|x) \log \frac{O_{\star,u}(y_+|x)}{O_{\star,u'}(y_+|x)} && (194)\\
&\le \frac{2}{X'Y'} \sum_{x \in \mathcal{X}', y \in \mathcal{Y}'_-} (1+\epsilon) \log \frac{1+\epsilon}{1-\epsilon} + (1-\epsilon) \log \frac{1-\epsilon}{1+\epsilon} && (195)\\
&\le \frac{2}{X'Y'} \sum_{x \in \mathcal{X}', y \in \mathcal{Y}'_-} 8\epsilon^2 && (196)\\
&= 8\epsilon^2 && (197)
\end{aligned}$$

for $\epsilon < 1/2$. Therefore, for if $K \le \frac{\log|\mathcal{U}| - 2\log 2}{16\epsilon^2}$, we have

$$\begin{aligned}
\max_{u \in \mathcal{U}} \mathbb{E}_u\left[v_u(\pi_u^\star) - v_u(\hat{\pi})\right] &\ge \frac{\epsilon}{16}\left(1 - \frac{8K\epsilon^2 + \log 2}{\log|\mathcal{U}|}\right) && (198)\\
&\ge \frac{\epsilon}{32} && (199)
\end{aligned}$$

From the lower bound on the size of $\mathcal{U}$, we have that

$$K \le \frac{X'Y'}{1024\epsilon^2} \tag{200}$$

implies the above condition because

$$\begin{aligned}
K &\le \frac{X'Y'}{1024\epsilon^2} && (201)\\
&\le \frac{X'Y'/32 - 2\log 2}{16\epsilon^2} && (202)\\
&\le \frac{\log|\mathcal{U}| - 2\log 2}{16\epsilon^2} && (203)
\end{aligned}$$

as long as $X'Y' \ge 64 \cdot 2\log 2$. Since $Y' = Y - 2$ and $X' = \frac{X+1}{2}$, we have

$$X'Y' = \frac{(X+1)(Y-2)}{2} \ge \frac{(X+1)Y}{4} \ge 128\log 2$$

where the second inequality follows from the constraints on X and $Y \ge 6$.

E. Proof of Theorem 6.3

A core component of the analysis is a simulation lemma that bounds the difference in values of a policy on two different POMDPs via the total variation distance of their models.

Lemma E.1 (Simulation Lemma)**.** *Consider a POMDP model with transition matrix $\hat{T}$, emission matrix $\hat{O}$, and reward function r. Denote the value function and measure under this POMDP by $\hat{v}$ and $\hat{P}$ respectively. Then, for any history-dependent policy π,*

$$\|\hat{P}_\pi - P_\pi\|_1 \le \mathbb{E}_\pi \sum_{h\in[H]} \|O_\star(\cdot|x_h) - \hat{O}(\cdot|x_h)\|_1 + \|T_\star(\cdot|x_h,a_h) - \hat{T}(\cdot|x_h,a_h)\|_1.$$

Furthermore,

$$|v(\pi) - \hat{v}(\pi)| \le H\mathbb{E}_\pi \sum_{h\in[H]} \|O_\star(\cdot|x_h) - \hat{O}(\cdot|x_h)\|_1 + \|T_\star(\cdot|x_h,a_h) - \hat{T}(\cdot|x_h,a_h)\|_1,$$

where $\mathbb{E}_\pi$ denotes the expectation following policy π under the true model $\mathcal{M}$.

This observation, although crude sometimes,[4] is useful in this setting because it is possible to estimate the models in HOMDP, in contrast to the POMDP where faithful recovery of $O_\star$ and $T_\star$ is not always possible.

E.1. High-probability events

Similar to the tabular setting, we define the following events and later show that they each occur with high probability.

$$\mathcal{E}_\mathcal{T} = \left\{ \forall k \in [K'], T \in \mathcal{T}, \sum_{\ell\in[k],h\in[H]} \|T(\cdot|x_h^\ell,a_h^\ell) - T_\star(\cdot|x_h^\ell,a_h^\ell)\|_1^2 \le 8\log(K'|\mathcal{T}|/\delta) + 4\sum_{\ell,h}\log\frac{T_\star(\tilde{x}_h^\ell|x_h^\ell,a_h^\ell)}{T(\tilde{x}_h^\ell|x_h^\ell,a_h^\ell)} \right\}$$

$$\mathcal{E}_\Theta = \left\{ \forall k \in [K'], O \in \Theta, \sum_{\ell\in[k],h\in[H]} \|O(\cdot|x_h^\ell) - O_\star(\cdot|x_h^\ell)\|_1^2 \le 8\log(K'|\Theta|/\delta) + 4\sum_{\ell,h}\log\frac{O_\star(y_h^\ell|x_h^\ell)}{O(y_h^\ell|x_h^\ell)} \right\}$$

The intersection of the above two is defined as $\mathcal{E}_{\mathcal{T},\Theta} = \mathcal{E}_\mathcal{T} \cap \mathcal{E}_\Theta$. Finally, let $\mathcal{E}_{\text{Fre}}$ denote the event that,for all $k \in [K']$ and $h \in [H]$, with $\tilde{\pi}_\ell = \hat{\pi}_\ell \circ_h \text{Unif}(\mathcal{A})$,

$$\sum_{\ell\in[k-1]} \mathbb{E}_{\tilde{\pi}_\ell}\left[\|T_\star(\cdot|x_h) - \hat{T}_k(\cdot|x_h,a_h)\|_1^2 + \|O_\star(\cdot|x_h,a_h) - \hat{O}_k(\cdot|x_h)\|_1^2\right] \tag{204}$$

$$\le 32\log\left(K'H|\mathcal{T}\times\Theta|/\delta\right) + 2\sum_{\ell\in[k-1]} \|T_\star(\cdot|x_h^\ell,a_h^\ell) - \hat{T}_k(\cdot|x_h^\ell,a_h^\ell)\|_1^2 + \|O_\star(\cdot|x_h^\ell) - \hat{O}_k(\cdot|x_h^\ell)\|_1^2. \tag{205}$$

Lemma E.2. $P(\mathcal{E}_\mathcal{T}) \ge 1-\delta$.

Proof. See Appendix E.4. □

Lemma E.3. $P(\mathcal{E}_\Theta) \ge 1-\delta$.

Proof. The proof is identical to that of Lemma E.2, except that one replaces $\mathcal{T}$ with Θ, $T_\star$ with $O_\star$, and the sample pairs $\tilde{x}_h^\ell, x_h^\ell, a_h^\ell$ with y_h^ℓ, x_h^ℓ. □

Lemma E.4. $P(\mathcal{E}_{Fre}) \ge 1-\delta$.

[4] Indeed, jumping straight to total variation bounds can lead to worse sample complexity bounds in the tabular case, which is why we opt for a more refined α-vector analysis for Theorem 4.2

Proof. Fix $T \in \mathcal{T}$ and $O \in \Theta$. For convenience, define

$$\epsilon(x,a) = \|T_\star(\cdot|x,a) - T(\cdot|x,a)\|_1 + \|O_\star(\cdot|x) - O(\cdot|x)\|_1. \tag{206}$$

Then, consider the stochastic process given by

$$Z_\ell = \epsilon(x_h^\ell, a_h^\ell)$$

It is easy to see that $Z_\ell - \mathbb{E}_{\tilde{\pi}_\ell}[Z_\ell]$ is a martingale difference sequence with $|Z_\ell| \le 8 =: R$ since x_h^ℓ, a_h^ℓ are drawn from the exploration policy $\tilde{\pi}_\ell$. By Theorem 1 of (Beygelzimer et al., 2011), with probability at least $1-\delta$, for all $k \in [K]$,

$$\sum_{\ell\in[k]} \mathbb{E}_{\tilde{\pi}_\ell}[Z_\ell] - Z_\ell \le \frac{1}{2R}\sum_{\ell\in[k]} \operatorname*{var}_{\tilde{\pi}_\ell}(Z_\ell) + 2R\log(1/\delta) \tag{207}$$

where $\operatorname{var}_{\tilde{\pi}_\ell}(Z_\ell) = \mathbb{E}_{\tilde{\pi}_\ell}\left(Z_\ell - \mathbb{E}_{\tilde{\pi}_\ell}[Z_\ell]\right)^2$. Then, note that

$$\operatorname*{var}_{\tilde{\pi}_\ell}(Z_\ell) \le \mathbb{E}_{\tilde{\pi}_\ell}\left[Z_\ell^2\right] \le R\mathbb{E}_{\tilde{\pi}_\ell} Z_\ell$$

since $0 \le Z_\ell \le R$. Applying this inequality and then rearranging, we have

$$\sum_{\ell\in[k]} \mathbb{E}_{\tilde{\pi}_\ell}[Z_\ell] \le 2\sum_{\ell\in[k]} Z_\ell + 4R\log(1/\delta)$$

Applying the definition of Z_ℓ and taking the union bound for all $h \in [H]$, $k \in [K']$, $T \in \mathcal{T}$ and $O \in \Theta$ gives the result with $R = 8$. □

E.2. Consequences of concentration

Lemma E.5. *Assume that the events $\mathcal{E}_{\mathcal{T},\Theta}$ and $\mathcal{E}_{Fre}$ hold. For all $k \in [K']$, $T_\star \in \mathcal{T}_k$ and $O_\star \in \Theta_k$.*

Proof. Note that from $\mathcal{E}_{\mathcal{T},\Theta}$, it holds that for all $k \in [K']$ and $T \in \mathcal{T}$ and $O \in \Theta$,

$$\sum_{\ell\in[k-1],h} \log T(\tilde{x}_h^\ell|x_h^\ell, a_h^\ell) - 2\log(K'|\mathcal{T}|/\delta) \le \sum_{\ell\in[k-1],h} \log T_\star(\tilde{x}_h^\ell|x_h^\ell, a_h^\ell) \tag{208}$$

and

$$\sum_{\ell\in[k-1],h} \log O(y_h^\ell|x_h^\ell) - 2\log(K'|\Theta|/\delta) \le \sum_{\ell\in[k-1],h} \log O_\star(y_h^\ell|x_h^\ell) \tag{209}$$

Given the definitions of $\beta_{\mathcal{T}}$ and β_Θ, the result is immediate. □

Lemma E.6. *Assume that the events $\mathcal{E}_{\mathcal{T},\Theta}$ and $\mathcal{E}_{Fre}$ hold. Then, for all $k \in [K']$ and $h \in [H]$, with $\tilde{\pi}_\ell = \hat{\pi}_\ell \circ_h \mathrm{Unif}(\mathcal{A})$,*

$$\begin{aligned}&\sum_{\ell\in[k-1]} \mathbb{E}_{\tilde{\pi}_\ell}\left[\|T_\star(\cdot|x_h,a_h) - \hat{T}_k(\cdot|x_h,a_h)\|_1^2 + \|O_\star(\cdot|x_h) - \hat{O}_k(\cdot|x_h)\|_1^2\right]\\ &\le 32\log(K'H|\mathcal{T}\times\Theta|/\delta) + 16\left(\beta_\Theta + \beta_{\mathcal{T}}\right)\end{aligned}$$

Proof. From $\mathcal{E}_{\mathrm{Fre}}$. Fix $h \in [H]$ and $k \in [K']$. Then,

$$\sum_\ell \mathbb{E}_{\tilde{\pi}_\ell}\left[\|T_\star(\cdot|x_h,a_h) - \hat{T}_k(\cdot|x_h,a_h)\|_1^2 + \|O_\star(\cdot|x_h) - \hat{O}_k(\cdot|x_h)\|_1^2\right] \tag{210}$$

$$\le 32\log(K'H|\mathcal{T}\times\Theta|/\delta) + 2\sum_\ell \|T_\star(\cdot|x_h^\ell,a_h^\ell) - \hat{T}_k(\cdot|x_h^\ell,a_h^\ell)\|_1^2 + \|O_\star(\cdot|x_h^\ell) - \hat{O}_k(\cdot|x_h^\ell)\|_1^2 \tag{211}$$

Then, using $\mathcal{E}_{\mathcal{T},\Theta}$,

$$\sum_{\ell} \|O_\star(\cdot|x_h^\ell) - \hat{O}_k(\cdot|x_h^\ell)\|_1^2 \le \sum_{\ell,h} \|O_\star(\cdot|x_h^\ell) - \hat{O}_k(\cdot|x_h^\ell)\|_1^2 \tag{212}$$

$$\le 8\log(K'|\Theta|/\delta) + 4\sum_{\ell,h} \log \frac{O_\star(y_h^\ell|x_h^\ell)}{\hat{O}_k(y_h^\ell|x_h^\ell)} \tag{213}$$

$$\le 16\log(K'|\Theta|/\delta) \tag{214}$$

where the last inequality uses the fact that $\hat{O}_k \in \Theta_k$ and $\max_{O\in\Theta_k} \sum_{\ell,h} \log O(y_h^\ell|x_h^\ell) \ge \sum_{\ell,h} \log O_\star(y_h^\ell|x_h^\ell)$ since $O_\star \in \Theta_k$. The same can be done for $\hat{T}_k$ and $\mathcal{T}_k$. Then the prior display can be bounded as

$$\sum_{\ell} \mathbb{E}_{\tilde{\pi}_\ell}\left[\|T_\star(\cdot|x_h,a_h) - T(\cdot|x_h,a_h)\|_1^2 + \|O_\star(\cdot|x_h) - O(\cdot|x_h)\|_1^2\right] \tag{215}$$

$$\le 32\log(K'H|\mathcal{T}\times\Theta|/\delta) + 32\left(\log(K'|\Theta|/\delta) + \log(K'|\mathcal{T}|/\delta)\right) \tag{216}$$

$$= 32\log(K'H|\mathcal{T}\times\Theta|/\delta) + 16\left(\beta_\Theta + \beta_\mathcal{T}\right). \tag{217}$$

□

E.3. Final steps

The final steps of the proof follow a classic optimism analysis. We let $\hat{v}_k$ denote the value function of the POMDP under the model transition function $\hat{T}_k$ and emission function $\hat{O}_k$ (selected optimistically in the algorithm).

The instantaneous regret for $k \in [K']$ is bounded as

$$v(\pi^\star) - v(\hat{\pi}_k) \le \hat{v}_k(\hat{\pi}_k) - v(\hat{\pi}_k) \tag{218}$$

$$\le H\sum_h \mathbb{E}_{\hat{\pi}_k} \|T_\star(\cdot|x_h,a_h) - \hat{T}_k(\cdot|x_h,a_h)\|_1 + \|O_\star(\cdot|x_h) - \hat{O}_k(\cdot|x_h)\|_1 \tag{219}$$

where the second line follows from the POMDP Simulation Lemma (Lemma E.1). Next, leveraging the low rank MDP assumption, define the following quantities:

$$\epsilon(T,O,x,a) := \|T(\cdot|x,a) - T_\star(\cdot|x,a)\|_1 + \|O(\cdot|x) - O_\star(\cdot|x)\|_1 \tag{220}$$

$$U_h(\hat{\pi}_k) := \mathbb{E}_{x_{h-1},a_{h-1}\hat{\pi}_k}\left[\phi(x_{h-1},a_{h-1})\right] \tag{221}$$

$$W_h(T,O) = \int_{x_h} \psi_\star(x_h)\cdot \sup_{a_h} \epsilon(T,O,x_h,a_h)\cdot \mathrm{d}x_h \tag{222}$$

$$\Sigma_{k,h} := \lambda I + \sum_{\ell\in[k-1]} U_h(\hat{\pi}_\ell)U_h(\hat{\pi}_\ell)^\top \tag{223}$$

for some $\lambda > 0$ to be determined later. For $h-1=0$, we can take U_h to be a fixed indicator and W_h to also be an indicator with a non-zero value of $\mathbb{E}_{x_1\sim\rho}\sup_a \epsilon(T,O,x_1,a)$. The algorithm does not need to use the vector functions U_h or W_h or the covariance matrix $\Sigma_{k,h}$. Only the analysis uses them. Then, for a fixed $h\in[H]$, let $\tilde{\pi}_\ell = \hat{\pi}_\ell \circ_h \mathrm{Unif}(\mathcal{A})$ be the exploration policy used in round ℓ for timestep h. Then, letting τ_h denote the concatenation of $(\tau_{h-1}, a_{h-1}, y_h)$ as usual,

$$\mathbb{E}_{\hat{\pi}_k}\|T_\star(\cdot|x_h,a_h) - \hat{T}_k(\cdot|x_h,a_h)\|_1 + \|O_\star(\cdot|x_h) - \hat{O}_k(\cdot|x_h)\|_1 \tag{224}$$

$$= \mathbb{E}_{x_{h-1},a_{h-1},\tau_{h-1}\sim\hat{\pi}_k}\left\langle \phi_\star(x_{h-1},a_{h-1}), \int_{y_h,x_h,a_h} \psi_\star(x_h)O_\star(y_h|x_h)\hat{\pi}_k(a_h|\tau_h)\epsilon(\hat{T}_k,\hat{O}_k,x_h,a_h)\cdot \mathrm{d}(y_h,x_h,a_h)\right\rangle \tag{225}$$

$$\le \mathbb{E}_{x_{h-1},a_{h-1}\sim\hat{\pi}_k}\left\langle \phi_\star(x_{h-1},a_{h-1}), \int_{y_h,x_h} \psi_\star(x_h)O_\star(y_h|x_h)\sup_{a_h}\epsilon(\hat{T}_k,\hat{O}_k,x_h,a_h)\cdot \mathrm{d}(y_h,x_h)\right\rangle \tag{226}$$

$$\le \|U_h(\hat{\pi}_k)\|_{\Sigma_{k,h}^{-1}} \cdot \left\| \int_{x_h} \psi_\star(x_h)\sup_{a_h}\epsilon(\hat{T}_k,\hat{O}_k,x_h,a_h)\cdot \mathrm{d}x_h\right\|_{\Sigma_{k,h}} \tag{227}$$

$$= \|U_h(\hat{\pi}_k)\|_{\Sigma_{k,h}^{-1}} \cdot \|W_h(\hat{T}_k,\hat{O}_k)\|_{\Sigma_{k,h}}. \tag{228}$$

The first line uses the definition of the low-rank latent transition to decompose the expectation over elements at step h. The first inequality replaces the distribution over a_h with a sup, which is valid because $\phi_\star(x_{h-1}, a_{h-1})^\top \psi_\star(x_h) \geq 0$ is a probability. The second inequality applies the Cauchy-Schwarz inequality with the definition of U_h and the last line uses the definition of W_h. For the right-hand factor,

$$\|W_h(\hat{T}_k, \hat{O}_k)\|^2_{\Sigma_{k,h}} = \lambda \|W_h(\hat{T}_k, \hat{O}_k)\|_2^2 + \sum_{\ell \in [k-1]} \left\langle U_h(\hat{\pi}_\ell), W_h(\hat{T}_k, \hat{O}_k) \right\rangle^2 \tag{229}$$

Using the normalization condition on $\psi_\star$, we have

$$\lambda \|W_h(\hat{T}_k, \hat{O}_k)\|_2^2 \leq 8\lambda d \tag{230}$$

Also,

$$\sum_{\ell \in [k-1]} \left\langle U_h(\hat{\pi}_\ell), W_h(\hat{T}_k, \hat{O}_k) \right\rangle^2 \tag{231}$$

$$= \sum_{\ell \in [k-1]} \left(\mathbb{E}_{x_h \sim \hat{\pi}_\ell} \sup_{a_h} \epsilon(\hat{T}_k, \hat{O}_k, x_h, a_h) \right)^2 \tag{232}$$

$$\leq 2 \sum_{\ell \in [k-1]} \mathbb{E}_{x_h \sim \hat{\pi}_\ell} \sup_{a_h} \left[\|T_\star(\cdot|x_h, a_h) - \hat{T}_k(\cdot|x_h, a_h)\|_1^2 + \|O_\star(\cdot|x_h) - \hat{O}_k(\cdot|x_h)\|_1^2 \right] \tag{233}$$

$$\leq 2A \sum_{\ell \in [k-1]} \mathbb{E}_{x_h, a_h \sim \tilde{\pi}_\ell} \left[\|T_\star(\cdot|x_h, a_h) - \hat{T}_k(\cdot|x_h, a_h)\|_1^2 + \|O_\star(\cdot|x_h) - \hat{O}_k(\cdot|x_h)\|_1^2 \right]. \tag{234}$$

The first inequality above applies Jensen's inequality and the fact that $(a+b)^2 \leq 2(a^2+b^2)$. The second inequality upper bounds the $\sup_{a_h}$ with a sum to convert the expression to a uniform distribution over a_h, which is exactly the distribution under the exploration policy $\tilde{\pi}_\ell$. Therefore, leveraging Lemma E.6 under $\mathcal{E}_\Theta$, $\mathcal{E}_\mathcal{T}$, and $\mathcal{E}_{\text{Fre}}$,

$$\|W_h(\hat{T}_k, \hat{O}_k)\|^2_{\Sigma_{k,h}} \leq 8\lambda d + 64A \left(\log\left(K'H|\mathcal{T} \times \Theta|/\delta\right) + \beta_\mathcal{T} + \beta_\Theta \right) \tag{235}$$

Then,

$$\mathbb{E}_{\hat{\pi}_k} \|T_\star(\cdot|x_h, a_h) - \hat{T}_k(\cdot|x_h, a_h)\|_1 + \|O_\star(\cdot|x_h) - \hat{O}_k(\cdot|x_h)\|_1 \tag{236}$$

$$\leq \min\left\{4, \sqrt{64A} \|U_h(\hat{\pi}_k)\|_{\Sigma^{-1}_{k,h}} \sqrt{\lambda d + \beta_\mathcal{T} + \beta_\Theta + \log(K'H|\mathcal{T} \times \Theta|/\delta)}\right\} \tag{237}$$

Let $\beta_\lambda := \lambda d + \beta_\mathcal{T} + \beta_\Theta + \log(K'H|\mathcal{T} \times \Theta|/\delta)$ for shorthand. Then, the total sub-optimality of the proposed policies $\hat{\pi}_1, \ldots, \hat{\pi}_K$ is bounded as

$$\sum_{k \in [K']} v(\pi^\star) - v(\hat{\pi}_k) \leq H \sum_{k,h} \left(4 \wedge \|U_h(\hat{\pi}_k)\|_{\Sigma^{-1}_{k,h}} \cdot \sqrt{64\beta_\lambda A}\right) \tag{238}$$

$$\leq H \sum_h \sqrt{64\beta_\lambda AK \sum_k \left(1 \wedge \|U_h(\hat{\pi}_k)\|^2_{\Sigma^{-1}_{k,h}}\right)} \tag{239}$$

$$\leq H^2 \sqrt{64\beta_{1/d} AKd \log(1+K')} \tag{240}$$

where the last inequality applies the elliptical potential lemma (Lemma F.8) with the setting $\lambda = 1/d$.

E.4. Proof of Lemma E.2

Proof. The proof follows a similar approach as Agarwal et al. (2020). Fix $k \in [K']$. In contrast to the the rest of the paper, in this proof only we use $\bar{\tau}^\ell$ to denote the data collected, since these differ from the histories in the usual sense as a result of the exploration happening over H rounds per epoch. In particular, we define $\bar{\tau}^\ell = \left\{x_h^\ell, a_h^\ell, y_h^\ell, \tilde{x}_h^\ell\right\}_{h=\in[H]}$ where we recall in the algorithm that $\tilde{x}_h^\ell$ is the next-state sampled by the exploration policy for step h.

Given the full trajectories $\bar{\tau}^1, \ldots, \bar{\tau}^k$, define the tangent dataset sampled independently as $\hat{x}^\ell_{h+1} \sim T_\star(\cdot|x^\ell_h, a^\ell_h)$ for $\ell \in [k]$, which has a distribution completely determined by the history. Let $l : \mathcal{X}^2 \times \mathcal{A} \to \mathbb{R}$ be an arbitrary measurable loss function. Define $L = \sum_{\ell,h} l(x^\ell_h, a^\ell_h, \tilde{x}^\ell_h)$ and $\hat{L} = \sum_{\ell,h} l(x^\ell_h, a^\ell_h, \hat{x}^\ell_h)$. For convenience, let us define $\bar{\boldsymbol{\tau}}^\ell_h$ to be the concatenated sequence $(\bar{\tau}^{1:\ell-1}, \bar{\tau}^\ell_h, a^\ell_h)$.

Consider the function:

$$\begin{aligned}\exp\left(L - \log \mathbb{E}\left[\exp(\hat{L}) \mid \bar{\boldsymbol{\tau}}^k_H\right]\right) &= \frac{\exp(L)}{\mathbb{E}\left[\exp(\hat{L}) \mid \bar{\boldsymbol{\tau}}^k_H\right]} \\ &= \frac{\exp(L)}{\prod_{\ell,h} \mathbb{E}\left[\exp l(x^\ell_h, a^\ell_h, \hat{x}^\ell_h) \mid \bar{\boldsymbol{\tau}}^\ell_h\right]}\end{aligned}$$

where the second equality follows from the fact that the tangent observations are independent given the history of latent states. Then,

$$\begin{aligned}\mathbb{E}\left[\exp\left(L - \log \mathbb{E}\left[\exp(\hat{L}) \mid \bar{\boldsymbol{\tau}}^k_h\right]\right)\right] &= \mathbb{E}\left[\frac{\exp(L)}{\prod_{\ell,h} \mathbb{E}\left[\exp l(x^\ell_h, a^\ell_h, \hat{x}^\ell_h) \mid \bar{\boldsymbol{\tau}}^\ell_h\right]}\right] && (241)\\ &= \mathbb{E}\left[\frac{\prod_{\ell,h} \exp l(x^\ell_h, a^\ell_h, \tilde{x}^\ell_h)}{\prod_{\ell,h} \mathbb{E}\left[\exp l(x^\ell_h, a^\ell_h, \hat{x}^\ell_h) \mid \bar{\boldsymbol{\tau}}^\ell_h\right]}\right] && (242)\\ &= \mathbb{E}\left[\frac{\prod_{\ell\in[k], h\in[H-1]} \exp\left(l(x^\ell_h, a^\ell_h, \tilde{x}^\ell_h)\right)}{\prod_{\ell\in[k], h\in[H-1]} \mathbb{E}\left[\exp l(x^\ell_h, a^\ell_h, \hat{x}^\ell_{h+1}) \mid \bar{\boldsymbol{\tau}}^\ell_h\right]} \cdot \frac{\mathbb{E}\left[\exp\left(l(x^k_H, a^k_H, \tilde{x}^k_H)\right) \mid \bar{\boldsymbol{\tau}}^k_H\right]}{\mathbb{E}\left[\exp\left(l(x^k_H, a^k_H, \hat{x}^k_H)\right) \mid \bar{\boldsymbol{\tau}}^k_H\right]}\right] && (243)\\ &= \mathbb{E}\left[\frac{\prod_{\ell\in[k], h\in[H-1]} \exp\left(l(x^\ell_h, a^\ell_h, \tilde{x}^\ell_{h+1})\right)}{\prod_{\ell\in[k], h\in[H-1]} \mathbb{E}\left[\exp l(x^\ell_h, a^\ell_h, \hat{x}^\ell_{h+1}) \mid \bar{\boldsymbol{\tau}}^\ell_h\right]}\right] && (244)\\ &= \ldots && (245)\\ &= 1 && (246)\end{aligned}$$

where the cancellation is repeated for all $\ell \in [k]$ and $h \in [H]$. Applying Markov's inequality ensures that

$$P\left(L - \log \mathbb{E}\left[\exp(\hat{L}) \mid \bar{\boldsymbol{\tau}}^k_H\right] \geq z\right) \leq \frac{\mathbb{E}\left[\exp\left(L - \log \mathbb{E}\left[\exp(\hat{L}) \mid \bar{\boldsymbol{\tau}}^k_H\right]\right)\right]}{e^z}. \tag{247}$$

Taking $z = \log(1/\delta)$ guarantees that

$$L - \log \mathbb{E}\left[\exp(\hat{L}) \mid \bar{\boldsymbol{\tau}}^k_H\right] \leq \log(1/\delta) \tag{248}$$

with probability at least $1 - \delta$.

We will define the loss function as $\ell(x, a, x') = \log \sqrt{\frac{T(x'|x,a)}{T_\star(x'|x,a)}}$ for some $T \in \mathcal{T}$. One can then relate the above loss function to the total variation distance:

$$\begin{aligned}\sum_{\ell\in[k], h\in[H]} \|T_\star(\cdot|x^\ell_h, a^\ell_h) - T(\cdot|x^\ell_h, a^\ell_h)\|_1^2 &= \sum_{\ell,h}\left(\int_{x'\in\mathcal{X}} |T_\star(x'|x^\ell_h, a^\ell_h) - T(x'|x^\ell_h, a^\ell_h)| \,\mathrm{d}x'\right)^2 && (249)\\ &\leq 4\sum_{\ell,h}\int_{x'}\left(\sqrt{T_\star(x'|x^\ell_h, a^\ell_h)} - \sqrt{T(x'|x^\ell_h, a^\ell_h)}\right)^2 \mathrm{d}x' && (250)\\ &= 8\sum_{\ell,h} \mathbb{E}_{x'\sim T_\star(\cdot|x^\ell_h, a^\ell_h)}\left[1 - \sqrt{T(x'|x^\ell_h, a^\ell_h)/T_\star(x'|x^\ell_h, a^\ell_h)}\right] && (251)\\ &\leq -8\sum_{\ell,h} \log \mathbb{E}_{x'\sim T_\star(\cdot|x^\ell_h, a^\ell_h)}\sqrt{\frac{T(x'|x^\ell_h, a^\ell_h)}{T_\star(x'|x^\ell_h, a^\ell_h)}} && (252)\end{aligned}$$

where the first inequality is from Cauchy-Schwarz and the second uses the fact that $\log(1+a) \le a$ for $a > -1$. Observe that we have

$$-\log \mathbb{E}\left[\exp(\hat{L}) \mid \bar{\boldsymbol{\tau}}_H^k\right] = -\log \mathbb{E}\left[\prod_{\ell,h} \exp\left(\log\left(\sqrt{\frac{T(\hat{x}_{h+1}^\ell | x_h^\ell, a_h^\ell)}{T_\star(\hat{x}_{h+1}^\ell | x_h^\ell, a_h^\ell)}}\right)\right) \mid \bar{\boldsymbol{\tau}}_H^k\right] \tag{253}$$

$$= -\sum_{\ell,h} \log\left(\mathbb{E}_{x' \sim T_\star(\cdot|x_h^\ell, a_h^\ell)} \sqrt{\frac{T(x'|x_h^\ell, a_h^\ell)}{T_\star(x'|x_h^\ell, a_h^\ell)}}\right) \tag{254}$$

Combining this with the concentration inequality from earlier, we conclude that

$$\sum_{\ell,h} \|T_\star(\cdot|x_h^\ell, a_h^\ell) - T(\cdot|x_h^\ell, a_h^\ell)\|_1^2 \le -8 \log \mathbb{E}\left[\exp(\hat{L}) \mid \mathbf{x}_H^k\right] \tag{255}$$

$$\le -8\left(L - \log(1/\delta)\right) \tag{256}$$

$$= -8\left(\frac{1}{2}\sum_{\ell,h} \log\left(\frac{T(x_{h+1}^\ell | x_h^\ell, a_h^\ell)}{T_\star(x_{h+1}^\ell | x_h^\ell, a_h^\ell)}\right) - \log(1/\delta)\right) \tag{257}$$

with probability at least $1-\delta$ for a fixed T. Taking the union bound over all $T \in \mathcal{T}$ and all $k \in [K']$,

$$\sum_{\ell \in [k], h \in [H]} \|T_\star(\cdot|x_h^\ell, a_h^\ell) - T(\cdot|x_h^\ell, a_h^\ell)\|_1^2 \le 8\log(K'|\mathcal{T}|/\delta) + 4\sum_{\ell,h} \log \frac{T_\star(x_{h+1}^\ell | x_h^\ell, a_h^\ell)}{T(x_{h+1}^\ell, x_h^\ell, a_h^\ell)} \tag{258}$$

with probability at least $1-\delta$. □

F. Auxiliary Lemmas

Here state and prove a number of helpful auxiliary results for the main theorems.

F.1. Simulation lemma

Lemma E.1 (Simulation Lemma)**.** *Consider a POMDP model with transition matrix $\hat{T}$, emission matrix $\hat{O}$, and reward function r. Denote the value function and measure under this POMDP by $\hat{v}$ and $\hat{P}$ respectively. Then, for any history-dependent policy π,*

$$\|\hat{P}_\pi - P_\pi\|_1 \le \mathbb{E}_\pi \sum_{h \in [H]} \|O_\star(\cdot|x_h) - \hat{O}(\cdot|x_h)\|_1 + \|T_\star(\cdot|x_h, a_h) - \hat{T}(\cdot|x_h, a_h)\|_1.$$

Furthermore,

$$|v(\pi) - \hat{v}(\pi)| \le H\mathbb{E}_\pi \sum_{h \in [H]} \|O_\star(\cdot|x_h) - \hat{O}(\cdot|x_h)\|_1 + \|T_\star(\cdot|x_h, a_h) - \hat{T}(\cdot|x_h, a_h)\|_1,$$

where $\mathbb{E}_\pi$ denotes the expectation following policy π under the true model $\mathcal{M}$.

Proof. Note that

$$|v(\pi) - \hat{v}(\pi)| \le H\|P_\pi - \hat{P}_\pi\|_1$$

where P_π denotes the measure over trajectories $\bar{\tau} = (x_1, y_1, a_1, \ldots, x_H, y_H, a_H)$ including the latent states under the true model with T and O and policy π. Similarly $\hat{P}_\pi$ denotes the same measure but under the model with $\hat{T}$ and $\hat{O}$. With notation slightly abused, we also use $P_\pi(\bar{\tau})$ to denote the density. This density decomposes as

$$P_\pi(\bar{\tau}) = \rho(x_1)\left(\prod_{h \in [H-1]} T_\star(x_{h+1}|x_h, a_h) O_\star(y_h|x_h)\pi(a_h|\tau_h)\right) O_\star(y_H|x_H)\pi(a_H|\tau_H),$$

where we recall that $\tau_h = (y_{1:h}, a_{1:h-1})$ is the partial history. $\hat{P}_\pi$ is analogously defined. Consider for now a fixed $\bar{\tau}$. To bound the total variation distance, we are interested in bounding the differences between $P_\pi(\bar{\tau})$ and $\hat{P}_\pi(\bar{\tau})$. For shorthand, we will define the following quantities:

$$B_h = \rho(x_1)O_\star(y_1|x_1) \prod_{t\in[h-1]} T_\star(x_{t+1}|x_t, a_t)O_\star(y_{t+1}|x_{t+1}),$$
$$\bar{\pi}_h = \prod_{t\in[h]} \pi(a_t|\tau_t).$$

Note that we have the following recursion:

$$B_h = O_\star(y_h|x_h)T_\star(x_h|x_{h-1}, a_{h-1})B_{h-1} \tag{259}$$

where $B_1 = \rho(x_1)O_\star(y_1|x_1)$. We also define $\hat{B}_h$ analogously with $\hat{O}$ and $\hat{T}$. To prove Lemma E.1, we will recursively apply the following bound.

Lemma F.1. *The following inequality holds:*

$$\int_{x_h,y_h,a_h} \bar{\pi}_h|B_h - \hat{B}_h| \cdot \mathrm{d}(x_h, y_h, a_h) \le \bar{\pi}_{h-1} \int_{x_h} T_\star(x_h|x_{h-1}, a_{h-1})B_{h-1} \cdot \|O_\star(\cdot|x_h) - \hat{O}(\cdot|x_h)\|_1 \cdot \mathrm{d}x_h \tag{260}$$
$$+ \bar{\pi}_{h-1}B_{h-1} \cdot \|T_\star(\cdot|x_{h-1}, a_{h-1}) - \hat{T}(\cdot|x_{h-1}, a_{h-1})\|_1 \tag{261}$$
$$+ \bar{\pi}_{h-1}|B_{h-1} - \hat{B}_{h-1}| \tag{262}$$

Summing over all possible latent-augmented trajectories, this implies that we have

$$\int_{\bar{\tau}} |P_\pi(\bar{\tau}) - P_\pi(\bar{\tau})| \cdot \mathrm{d}\bar{\tau} = \int_{\bar{\tau}} \bar{\pi}_H|B_H - \hat{B}_H| \cdot \mathrm{d}\bar{\tau} \tag{263}$$
$$\le \sum_h \int_{x_h} P_\pi(x_h) \cdot \|O_\star(\cdot|x_h) - \hat{O}(\cdot|x_h)\|_1 \cdot \mathrm{d}x_h \tag{264}$$
$$+ \sum_{h\in[H-1]} \int_{x_h,a_h} P_\pi(x_h, a_h) \cdot \|T_\star(\cdot|x_h, a_h) - \hat{T}(\cdot|x_h, a_h)\|_1 \tag{265}$$
$$= \mathbb{E}_\pi \sum_h \|O_\star(\cdot|x_h) - \hat{O}(\cdot|x_h)\|_1 + \mathbb{E}_\pi \sum_{h\in[H-1]} \|T_\star(\cdot|x_h, a_h) - \hat{T}(\cdot|x_h, a_h)\|_1. \tag{266}$$

□

F.1.1. Proof of Lemma F.1

Lemma F.1. *The following inequality holds:*

$$\int_{x_h,y_h,a_h} \bar{\pi}_h|B_h - \hat{B}_h| \cdot \mathrm{d}(x_h, y_h, a_h) \le \bar{\pi}_{h-1} \int_{x_h} T_\star(x_h|x_{h-1}, a_{h-1})B_{h-1} \cdot \|O_\star(\cdot|x_h) - \hat{O}(\cdot|x_h)\|_1 \cdot \mathrm{d}x_h \tag{260}$$
$$+ \bar{\pi}_{h-1}B_{h-1} \cdot \|T_\star(\cdot|x_{h-1}, a_{h-1}) - \hat{T}(\cdot|x_{h-1}, a_{h-1})\|_1 \tag{261}$$
$$+ \bar{\pi}_{h-1}|B_{h-1} - \hat{B}_{h-1}| \tag{262}$$

Proof. We first average out a_h and then apply the triangle inequality to bound the quantity in terms of the difference in

emission matrices $\|O(\cdot|x_h) - \hat{O}(\cdot|x_h)\|_1$:

$$\int_{x_h,y_h,a_h} \bar{\pi}_h |B_h - \hat{B}_h| \cdot \mathrm{d}(x_h, y_h, a_h) \tag{267}$$

$$= \bar{\pi}_{h-1} \int_{x_h,y_h} |B_h - \hat{B}_h| \cdot \mathrm{d}(x_h, y_h) \tag{268}$$

$$= \bar{\pi}_{h-1} \int_{x_h,y_h} |T_\star(x_h|x_{h-1}, a_{h-1}) O_\star(y_h|x_h) B_{h-1} - \hat{T}(x_h|x_{h-1}, a_{h-1}) \hat{O}(y_h|x_h) \hat{B}_{h-1}| \cdot \mathrm{d}(x_h, y_h) \tag{269}$$

$$\leq \bar{\pi}_{h-1} \int_{x_h,y_h} T_\star(x_h|x_{h-1}, a_{h-1}) B_{h-1} \cdot |O_\star(y_h|x_h) - \hat{O}(y_h|x_h)| \cdot \mathrm{d}(x_h, y_h) \tag{270}$$

$$+ \bar{\pi}_{h-1} \int_{x_h,y_h} \hat{O}(y_h|x_h) |T_\star(x_h|x_{h-1}, a_{h-1}) B_{h-1} - \hat{T}(x_h|x_{h-1}, a_{h-1}) \hat{B}_{h-1}| \cdot \mathrm{d}(x_h, y_h) \tag{271}$$

$$= \underbrace{\bar{\pi}_{h-1} \int_{x_h} T_\star(x_h|x_{h-1}, a_{h-1}) B_{h-1} \cdot \|O_\star(\cdot|x_h) - \hat{O}(\cdot|x_h)\|_1 \cdot \mathrm{d}(x_h)}_{\textbf{(I)}} \tag{272}$$

$$+ \bar{\pi}_{h-1} \int_{x_h} |T_\star(x_h|x_{h-1}, a_{h-1}) B_{h-1} - \hat{T}(x_h|x_{h-1}, a_{h-1}) \hat{B}_{h-1}| \cdot \mathrm{d}(x_h). \tag{273}$$

Now, we can also apply the triangle inequality to the last term on the right side to bound the quantity in terms of the difference in transition matrices:

$$\int_{x_h,y_h,a_h} \bar{\pi}_h |B_h - \hat{B}_h| \cdot \mathrm{d}(x_h, y_h, a_h) \leq \textbf{(I)} + \bar{\pi}_{h-1} B_{h-1} \int_{x_h} |T_\star(x_h|x_{h-1}, a_{h-1}) - \hat{T}(x_h|x_{h-1}, a_{h-1})| \cdot \mathrm{d}(x_h) \tag{274}$$

$$+ \bar{\pi}_{h-1} \int_{x_h} \hat{T}(x_h|x_{h-1}, a_{h-1}) |B_{h-1} - \hat{B}_{h-1}| \cdot \mathrm{d}(x_h) \tag{275}$$

$$\leq \textbf{(I)} + \bar{\pi}_{h-1} B_{h-1} \cdot \|T(\cdot|x_{h-1}, a_{h-1}) - \hat{T}(\cdot|x_{h-1}, a_{h-1})\|_1 \tag{276}$$

$$+ \bar{\pi}_{h-1} |B_{h-1} - \hat{B}_{h-1}|. \tag{277}$$

This concludes the proof. □

F.2. Concentration inequalities

Lemma F.2 (Hoeffding's inequality)**.** *Let $Z_1, \ldots, Z_n$ be a sequence of independent random variables with $Z_i \in [a, b]$ for all i for $-\infty < a \leq b < \infty$. Then*

$$P\left(\frac{1}{n}\sum_i Z_i - \mathbb{E}[Z_i] \geq (b-a)\sqrt{\frac{\log(1/\delta)}{n}}\right) \leq \delta.$$

Lemma F.3 (Bernstein's inequality)**.** *Let $Z_1, \ldots, Z_n$ be a sequence of independent random variables with $Z_i \in [0, 1]$ and variance $\mathrm{var}(Z_i) = \sigma^2$ and mean $\mathbb{E}[Z_i] = \mu$ for all i. Then, with probability at least $1 - \delta$,*

$$\frac{1}{n}\sum_i Z_i - \mathbb{E}[Z_i] \leq \frac{\log(1/\delta)}{3n} + \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}}.$$

Furthermore, this implies that, for all $c \geq 1$,

$$\begin{aligned}\frac{1}{n}\sum_i Z_i - \mu &\leq \frac{\log(1/\delta)}{3n} + \sqrt{\frac{2\mu \log(1/\delta)}{n}} \\ &\leq \frac{\mu}{2c} + \frac{2c \log(1/\delta)}{n}.\end{aligned}$$

Proof. The first statement is simply the original statement of Bernstein's inequality. The second uses the fact that $\sigma^2 \leq \mu$ for variables in $[0, 1]$. The last one uses the AM-GM inequality. □

Lemma F.4 (Azuma-Hoeffding). *Let $Z_1, \ldots, Z_n$ be a martingale difference sequence with $|Z_i| \leq G$ for all i. Then, with probability at least $1 - \delta$,*

$$\sum_i Z_i \leq 4G\sqrt{n \log(1/\delta)}. \tag{278}$$

F.3. Pigeonhole lemmas

Lemma F.5 (Pigeonhole Principle). *The following inequalities hold:*

$$\sum_{k\in[K],h\in[H]} \sqrt{\frac{1}{\max\{1, n_k(x_h^k)\}}} \leq HX + 3\sqrt{HXK} \tag{279}$$

and

$$\sum_{k\in[K],h\in[H]} \sqrt{\frac{1}{\max\{1, n_k(x_h^k, a_h^k)\}}} \leq HXA + 3\sqrt{HXAK}. \tag{280}$$

Proof. We prove only the first as the second is equivalent up to summations over the actions. Note that

$$\sum_{k\in[K],h\in[H]} \sqrt{\frac{1}{\max\{1, n_k(x_h^k)\}}} = \sum_x \sum_{k=1}^K \frac{m_k(x)}{\sqrt{\max\{1, n_k(x)\}}} \tag{281}$$

$$\leq XH + \sum_x \sum_{k=1}^K \frac{m_k(x)}{\sqrt{\max\{H, n_k(x)\}}}, \tag{282}$$

where $m_k(x) = \sum_{h\in[H]} \mathbf{1}\{x_h^k = x\}$ counts the number of occurrences of x in a single round k. The inequality uses the fact that, for any x, the summand can contribute at most H to the sum (because $m_k(x)$ is bounded by H) before $n_k(x)$ has value at least H. Now we use Lemma F.6 (which is adapted from Lemma 19 of (Auer et al., 2008)) to bound the second term:

$$\sum_{k\in[K],h\in[H]} \sqrt{\frac{1}{\max\{H, n_k(x_h^k)\}}} \leq XH + 3\sum_x \sqrt{n_K(x)} \tag{283}$$

$$\leq XH + 3\sqrt{HXK}, \tag{284}$$

where the last line follows from the Cauchy-Schwarz inequality along with the fact that $\sum_x n_K(x) = KH$.

□

Lemma F.6 (Adapted from Auer et al. (2008)). *Let $z_1, \ldots, z_n \in [0, H]$ be an arbitrary sequence and let $Z_k = \max\{H, \sum_{k=1}^k z_k\}$. Then,*

$$\sum_{k\in[n]} \frac{z_k}{\sqrt{Z_{k-1}}} \leq 3\sqrt{Z_n}. \tag{285}$$

Proof. Consider the case where $n = 1$. Then, $Z_0 = H$. Furthermore,

$$\sum_{k\in[n]} \frac{z_k}{\sqrt{Z_k - 1}} = \frac{z_1}{\sqrt{H}} \leq \sqrt{H} \leq 3\sqrt{Z_1} \tag{286}$$

By induction on the base case, we have

$$\sum_{k\in[n]} \frac{z_k}{\sqrt{Z_{k-1}}} = 3\sqrt{Z_{n-1}} + \frac{z_n}{\sqrt{Z_{n-1}}} \tag{287}$$

$$\left(3\sqrt{Z_{n-1}} + \frac{z_n}{\sqrt{Z_{n-1}}}\right)^2 = 9Z_{n-1} + 6z_n + \frac{z_n^2}{Z_{n-1}} \tag{288}$$

$$\leq 9Z_{n-1} + 7z_n \tag{289}$$

$$\leq 9Z_n \tag{290}$$

Therefore, by taking the square root,

$$\sum_{k\in[n]} \frac{z_k}{\sqrt{Z_{k-1}}} \leq 3\sqrt{Z_n} \tag{291}$$

□

Lemma F.7. *The following inequality holds:*

$$\sum_{k\in[K],h\in[H]} \frac{1}{\max\left\{1, n_k(x_h^k, a_h^k)\right\}} \leq HXA(1+\log K). \tag{292}$$

Proof. First note that since $n_k(x,a)$ is updated each episode, we immediately have

$$\sum_{k\in[K],h\in[H]} \frac{1}{\max\left\{1, n_k(x_h^k)\right\}} \leq HXA \sum_{i=1}^{\lceil K/XA\rceil} \frac{1}{i} \tag{293}$$

$$\leq HXA \sum_{i=1}^{K} \frac{1}{i} \tag{294}$$

since we assume that $X, A \geq 1$. Then,

$$\sum_{i\in[K]} \frac{1}{i} \leq 1 + \int_1^K \frac{dx}{x} \tag{295}$$

$$= 1 + \log K. \tag{296}$$

□

Lemma F.8 (Lattimore & Szepesvári (2020))**.** *Let* $\Sigma_k = \lambda I + \sum_{\ell\in[k-1]} \phi_\ell \phi_\ell^\top$ *and* $\|\phi_\ell\| \leq 1$ *uniformly. Then,*

$$\sum_{k\in[K]} \left(1 \wedge \|\phi_k\|^2_{\Sigma_k^{-1}}\right) \leq 2d \log\left(\frac{d\lambda + K}{d\lambda}\right). \tag{297}$$