Prospective Side Information for Latent MDPs

Jeongyeol Kwon [1] **Yonathan Efroni** [2] **Shie Mannor** [3 4] **Constantine Caramanis** [5]

Abstract

In many interactive decision-making problems, there is contextual side information that remains fixed within the course of an interaction. This problem has been studied quite extensively under the assumption the context is fully observed, as well as in the opposing limit when the context is unobserved, a special type of POMDP also referred to as a Latent MDP (LMDP). In this work, we consider a class of decision problems that interpolates between the settings, namely, between the case the context is fully observed, and the case the context is unobserved. We refer to this class of decision problems as *LMDPs with prospective side information*. In such an environment an agent receives additional, weakly revealing, information on the latent context at the beginning of each episode. We show that, surprisingly, this problem is not captured by contemporary POMDP settings and is not solved by RL algorithms designed for partially observed environments. We then establish that any sample efficient algorithm must suffer at least $\Omega(K^{2/3})$-regret, as opposed to standard $\Omega(\sqrt{K})$ lower bounds. We design an algorithm with a matching upper bound that depends only polynomially on the problem parameters. This establishes exponential improvement in the sample complexity relatively to the existing LMDP lower bound, when prospective information is not given (Kwon et al., 2021).

1. Introduction

"If in the first act you have hung a pistol on the wall, then in the following one it should be fired." A. Chekhov famously stated. From a decision making perspective this idea serves as a reminder that in many environments an observer is being presented with side information at the beginning of an interaction that will be of value only in later use. Mathematically, this can be modelled as a contextual decision problem with initial side information. Such natural problem has been studied for the fully observed, Markovian setting, when an agent has access to the contextual side information (Jiang et al., 2017; Modi et al., 2018; Sun et al., 2019). Such a problem was additionally studied in the partially observed setting, when the contextual information exists but is not observable by the agent (Chadès et al., 2012; Hallak et al., 2015; Brunskill & Li, 2013; Chatterjee et al., 2020; Steimle et al., 2018; Kwon et al., 2021), a setting that was also referred as Latent MDPs (LMDPs) in prior literature.

An LMDP can model real world problems where there exists an unobserved latent context, *e.g.,* in dialogue, recommender or in healthcare systems, when complete information on a user or patient is not given, yet, each user remains fixed within each episodic interaction. Recently, Kwon et al. (2021; 2023) derived exponential worst-case lower bounds in the number of contexts for this subclass of POMDPs. This implies that, in general, near optimal policy of an LMDP cannot be learned efficiently when the number of latent context is large.

In this work, we study a natural sub-class of LMDPs in which weakly revealing information on the latent context is given to an agent at *the beginning of the interaction*. That can be thought of as an intermediate regime between the case the contextual side information is given to the case it is hidden and latent, as in general LMDPS. We refer to this class of environments as LMDP with Prospective Side Information, or as LMDP-Ψ. We introduce this class of problems, study lower bounds and matching upper bounds for this class, and show this setting can be learned efficiently, unlike general LMDPs.

Motivating Example. Consider a navigation task where the goal location is randomly selected at the beginning of an episode from a small set of goal locations. Without further assumptions, also an optimal policy of such POMDP will perform poorly, and may not reach the goal state with high probability. A natural way to improve the performance of an agent is to supply it with additional hints about the goal

[1]Wisconsin Institute for Discovery, Wisconsin, USA [2]Meta AI, New York, USA [3]Electrical Engineering, Technion, Haifa, Israel [4]NVIDIA [5]Electrical and Computer Engineering, University of Texas at Austin, Texas, USA. Correspondence to: Jeongyeol Kwon <jeongyeol.kwon@wisc.edu>, Yonathan Efroni <jonathan.efroni@gmail.com>.

Proceedings of the 41^{st} International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

	MDP	α-Revealing POMDP	LMDP with α-Prospective SI	LMDP	POMDP
(UB)	$\sqrt{AK}$	$\mathrm{poly}(A, \alpha^{-1})\sqrt{K}$	$\mathrm{poly}(A, \alpha^{-1}, M)K^{2/3}$	Unknown	$A^H\sqrt{K}$
(LB)	$\sqrt{AK}$	$\mathrm{poly}(A, \alpha^{-1})\sqrt{K}$	$\min\left(A^{\Omega(M)}\sqrt{K}, \mathrm{poly}(A, \alpha^{-1}, M)K^{2/3}\right)$	$A^{\Omega(M)}\sqrt{K}$	$A^{\Omega(H)}\sqrt{K}$

Table 1: Known regret upper and lower bounds in different classes of POMDPs, ordered by their degree of difficulty from the simplest to hardest (left to right). Dependencies on other problem parameters are omitted (*e.g.,* S and H). The results and the setting introduced in this work are highlighted in green.

state, *e.g.,* which area the goal is located. Such a hint is given at the beginning of the interaction and remains fixed throughout the episode, yet, this hint does not fully specify the location of the goal state. Such an environment is an instance of the LMDP-Ψ class.

Our Contributions. The main contributions of this work are the following (see also Table 1). We introduce the LMDP-Ψ setting and study its sample complexity. Specifically, we study the problem of learning a near-optimal policy of an LMDP-Ψ when the prospective side information weakly reveals information, quantified by a parameter α, on the true latent state. We provide a $\mathrm{poly}(A, \alpha^{-1})K^{2/3}$ regret upper bound, by building upon the pure exploration scheme developed in Huang et al. (2023). Our upper bound does not suffer exponential dependence in the number of latent contexts as in an LMDP, namely, in the absence of prospective information Kwon et al. (2021; 2023). We also derive a lower bound of $\Omega\left(\frac{A}{\alpha^2\epsilon^2}K^{2/3}\right)$ to this problem, unlike the $K^{1/2}$ rate one may expect.

Technically speaking, our work builds upon recent algorithmic advancements for POMDPs (Liu et al., 2023; Uehara et al., 2022; Huang et al., 2023). However, proper application of these requires care. Perhaps our most surprising finding is that the LMDP-Ψ class is not contained within POMDP classes previously known to be efficiently learnable; hence, new results should be established for this class. This fact also puts forward natural questions on generalizations of the LMDP-Ψ setting which we leave to the future.

2. Preliminaries

An episodic LMDP is defined as follows:

Definition 2.1 (Latent MDP)**.** An LMDP instance consists of a tuple $\theta := \left(\{p_m\}_{m=1}^M, \{\mathbb{T}_m\}_{m=1}^M, \{\mathbb{O}_m\}_{m=1}^M\right)$, where M is the number of latent contexts; $\{p_m\}_{m=1}^M$ are the mixing weights, the probability latent context m is drawn at the beginning of an episode; $\mathbb{T}_m \in \mathbb{R}^{S\times S\times A}, \mathbb{O}_m \in \mathbb{R}^{|\mathcal{O}|\times S\times A}$ are the transition probabilities and instant observation distribution of m^{th} MDPs, *i.e.,* $\mathbb{T}_m(s'|s,a) := \mathbb{P}(s'|m,s,a)$ and $\mathbb{O}_m(o,s,a) := \mathbb{P}(o|m,s,a)$ for state $s \in \mathcal{S}$, next state $s' \in \mathcal{S}$, action $a \in \mathcal{A}$, instanteneous observation $o \in \mathcal{O}$, and latent context $m \in [M]$.

We assume that for all $o \in \mathcal{O}$, there is a known reward-decoding function $r : \mathcal{O} \to \mathbb{R}$, and each reward is bounded $|r(o)| \le 1$. To simplify the discussion, we assume that the set of LMDP instances Θ has finite (but exponentially large) cardinality $|\Theta|$. Similarly, we also assume that the observation space is discrete and finite:

Assumption 1 (Observation Space)**.** Each observation attains a value in the set $\mathcal{O}$ which has finite but could be arbitrarily large cardinality $|\mathcal{O}|$.

All claims made in this paper hold similarly for the continuous model class with a standard ϵ-discretization of Θ with the extra discretization error analysis similar to Liu et al. (2022) and continuous observations Liu et al. (2023).

At the beginning of every episode, a latent and unobserved context $m \in [M]$ is sampled from a mixing distribution $\{p_m\}_{m=1}^M$ and is fixed for H time steps. Without loss of generality, we assume that the system starts from time-step $t = 0$ at a fixed initial state s_{dummy} and transits to other states following the initial state distribution of the chosen MDP regardless of taken actions (and we always see a dummy observation o_{dummy}).

Prospective Side Information for LMDPs. In this work, we assume the LMDP is augmented with prospective side information. Prospective side information is an additional observation given *prior to the beginning of the episode* and remains fixed along an episode. Let $\mathcal{I}$ be the set of prospective side information values, and is assumed to be finite but may be arbitrarily large. Let $\mathbb{I} \in \mathbb{R}^{|\mathcal{I}|\times M}$ be a context dependent emission matrix, *i.e.,* $\mathbb{I}(\iota, m) := \mathbb{P}(\iota|m)$. Further, we assume it provides some hint on the identity of the true latent MDP. Formally, we assume the following weakly revealing condition:

Assumption 2 (α-Weakly Revealing Prospective Side Information)**.** For any two belief vectors $\bar{v}_1, \bar{v}_2 \in \Delta([M])$,

$$d_{\mathrm{TV}}\left(\mathbb{P}(\iota|\bar{v}_1), \mathbb{P}(\iota|\bar{v}_2)\right) \ge \frac{\alpha}{2}\|\bar{v}_1 - \bar{v}_2\|_1. \tag{1}$$

With these definitions at hand, we define the LMDP with Prospective Side Information, or LMDP-Ψ, An LMDP-Ψ is the tuple $\theta = (\mathbb{I}, \{p_m, \mathbb{T}_m, \mathbb{O}_m\}_{m=1}^M) \in \Theta$.

Accordingly, our goal is now to learn an ϵ-optimal policy from a larger class of policies $\Pi : \mathcal{I} \times (\mathcal{A} \times \mathcal{O} \times \mathcal{S})^* \to$

$\Delta(\mathcal{A})$ that exploits the prospective side information. An additional subclass which is useful to define is the class of *side information blind policies* $\Pi_{\texttt{blind}} : (\mathcal{A} \times \mathcal{O} \times \mathcal{S})^* \to \Delta(\mathcal{A})$, that does not exploit the prospective side information. Trivially $\Pi_{\texttt{blind}} \subset \Pi$ if $\mathcal{I}$ is non-empty. As we will see, the nature of the problem becomes different by the capacity of the policy class.

The optimal policy π^* is the optimal history-dependent policy that maximizes the expected cumulative reward

$$V^\star = \max_{\pi \in \Pi} V^\pi := \mathbb{E}^\pi \left[\sum_{t=1}^H r_t(o_t) \right],$$

where the expectation is taken over latent contexts and rewards generated by an LMDP instance, following policy π. We let $\pi^*_{\texttt{blind}}$ be the counterpart in the smaller policy class $\Pi_{\texttt{blind}}$.

Notation We use the symbol $\lesssim$ to mean that the inequality holds up to some absolute multiplicative constant. We use $\lesssim_{\texttt{P}}$ when it holds up to some *problem dependent* polynomial factors. To simplify notation, we occasionally denote pairwise quantities as $x_t := (s_t, a_t)$, $y_t := (o_t, s_{t+1})$. $\mathrm{Ber}(p)$ denotes a Bernoulli random variable with parameter $p \in [0, 1]$. For arbitrary full column-rank matrix M, $M^\dagger$ is a left-inverse of M such that $M^\dagger M = I$.

3. Related Work

The study of learning algorithms for LMDPs was initiated within the framework of long-horizon multitask RL (Taylor & Stone, 2009; Brunskill & Li, 2013; Hallak et al., 2015; Liu et al., 2016), where full information on the latent contexts is revealed for a long-enough episode. However, problems in which full information on the latent context is not revealed cannot be solved through this framework. Kwon et al. (2021) considered the sample complexity of learning a near-optimal policy for LMDPs without any assumptions. Unfortunately, their lower bound is exponential in the number of contexts, even when the transition dynamics are shared (Kwon et al., 2023; 2022). Hence, further investigation on the natural assumption for which LMDPs are efficiently learnable is required. To overcome the fundamental barriers in LMDPs, a few works have considered the assumption of giving true information in hindsight (Kwon et al., 2021; Zhou et al., 2022; Lee et al., 2023), as discussed earlier.

Another related work to our setting is the *multi-step* weakly revealing POMDP, where an agent must play sub-optimal actions to obtain weakly-revealing information (Golowich et al., 2022; Liu et al., 2023; Chen et al., 2023). In this setting, a similar lower bound of $K^{2/3}$ regret has been reported in Chen et al. (2023). While our lower bound construction is partially inspired by theirs, the LMDP-Ψ setting is different since we obtain the weakly-revealing information "for free" at the beginning of each episode. Hence, a priori, one may hope for an improved upper bound in the simpler LMDP-Ψ setting.

Lastly, in Kwon et al. (2021); Zhou et al. (2022); Lee et al. (2023) the authors studied a somewhat dual setting to the one we consider here: they assume the agent receives complete information on the latent context in hindsight. Unlike their work, we assume the revealing information is being given at the initial time step: this implies the agent can use the prospective information during the interaction. Further, we do not assume the prospective information is sufficient for deterministically decoding the latent state as in these works.

LMDP-Ψ is not a Weakly Revealing POMDP. The recent line of work on weakly revealing POMDPs (Liu et al., 2022; 2023; Uehara et al., 2022; Chen et al., 2022; 2023) is the most closely related to ours. Next, we elaborate on the differences between the settings. These highlight both the novelty and challenges in tackling the LMDP-Ψ problem.

- *Standard POMDP modeling assumptions are violated in the presence of prospective information.* For the LMDP-Ψ setting, the available observations between different time steps *are not independent, conditioned on the latent state*. Let the available observation at each time step be $\tilde{o}_t := (o_t, \iota)$, *i.e.*, a combination of the observation and the available initial prospective side information. Trivially, the common conditional independence on the latent state assumption for the observation generation process does not hold. It does not necessarily hold that $\mathbb{P}^\pi(\tilde{o}_t \mid s_t, m) \neq \mathbb{P}^\pi(\tilde{o}_t \mid s_t, m, \tilde{o}_{t-1})$: $\tilde{o}_{t-1}$ contains information on $\tilde{o}_t$ since the prospective information, ι, is fixed during an episode. That is, there is a non-trivial correlation between observations. Unlike LMDP-Ψ, in the common POMDP and the weakly revealing POMDP settings (Liu et al., 2022), the observation is independent of historical information conditioned on the latent state.

- *Regret guarantees are fundamentally different.* As depicted in Table 1, the regret lower bound for LMDP-Ψ, without the exponential on the number of latent contexts, is $\Omega(K^{2/3})$. Such a lower bound is fundamentally different than the $O(\sqrt{K})$ upper bound for weakly revealing POMDPs. This highlights a key difference between the settings established by our results.

4. Learning in LMDP-Ψ

In this section, we present our algorithmic results as well as lower bound analysis.

Algorithm 1 Regret Minimization within $\Pi_{\texttt{blind}}$

Initialize $\mathcal{D}^0 = \emptyset, \mathcal{C}^0 = \Theta$
for $k = 1...K$ **do**
  # Optimistic Policy Search
  Pick $(\theta^k, \pi^k) = \arg\max_{\theta \in \mathcal{C}^k, \pi \in \Pi_{\texttt{blind}}} V_{\theta^k}^{\pi^k}$
  Get $\tau^k = (s_1^k, a_1^k, ..., r_H^k), \iota^k$ by executing π^k
  # Confidence Set Construction
  $\mathcal{D}^k \leftarrow \mathcal{D}^{k-1} \cup \{(\iota^k, \tau^k, \pi^k)\}$ and update $\mathcal{C}^k$ using (6)
end for

4.1. Warm Up: $\sqrt{K}$-Regret within $\Pi_{\texttt{blind}}$

Consider the problem of learning a near-optimal policy only in the blind policy class $\Pi_{\texttt{blind}}$. Such a setting is equivalent to the one in which the prospective side information is provided in hindsight, and thus, the problem falls into the setting of well-conditioned PSR studied in Liu et al. (2023). To see this, define problem operators $B(o, s_{+1}|s, a) = \mathbb{I} \cdot \textbf{diag}([\mathbb{P}(o, s_{+1}|m, s, a)]_{m=1}^{M}) \cdot \mathbb{I}^\dagger$ and $b_0 = \mathbb{I}w$. We can easily verify that for any blind policy $\pi \in \Pi_{\texttt{blind}}$ and trajectory $\tau = (s_1, a_1, o_1, ..., s_H, a_H, o_H)$,

$$\mathbb{P}^\pi(\iota, \tau) = \boldsymbol{e}_\iota^\top \cdot \Pi_{t=1}^H B(o_t, s_{t+1}|s_t, a_t) \cdot b_0 \cdot \pi(\tau),$$

where $\pi(\tau) = \Pi_{h=1}^H \pi(a_h|s_1, ..., s_h)$. We define $s_{H+1} := \emptyset$ in the above expression. Let

$$\omega_t := (r_t, s_{t+1}, a_{t+1}, ..., r_H),$$
$$\psi(\omega_t, \iota|s_t, a_t)^\top := \boldsymbol{e}_\iota^\top \cdot \Pi_{h=t}^H B(o_h, s_{h+1}|s_h, a_h), \quad (2)$$

where ω_t is the future partial trajectory from time step t. With this, the system reparameterized by B and b_0 with the blind policy class is a well-conditioned PSR, as defined in Liu et al. (2023) (see their Condition 4.3), *i.e.,* for any $t \in [H]$ and any policy $\pi \in \Pi_{\texttt{blind}}$ it holds that

$$\max_{b:\|b\|_1=1} \sum_{\iota, \omega_t} \pi(\omega_t)|\psi(\omega_t, \iota|s_t, a_t)^\top b| \le \frac{M}{\alpha}. \quad (3)$$

With the above condition, since no extra tests are required to obtain ι, this allows us to apply the Optimistic-MLE (O-MLE) algorithm introduced in Liu et al. (2022) for regret minimization (see Algorithm 1).

We can follow the analysis of the optimistic-MLE approach for well-conditioned PSRs (Liu et al., 2023), yielding the following theorem:

Theorem 4.1. *Let π^*_{blind} be the optimal policy in Π_{blind} for the true environment θ^*. With probability greater than $1-\delta$, the regret of Algorithm 1 (with respect to the optimal blind policy) satisfies*

$$\sum_{k=1}^K V_{\theta^*}^{\pi^*_{blind}} - V_{\theta^*}^{\pi^k} \lesssim \frac{M^{3/2}H^2}{\alpha}\sqrt{SAK\log(|\Theta|/\delta)(\log K)}.$$

Note that the size of model class $|\Theta|$ is typically exponential in the number of free parameters that define the system, and we would hope to bound the regret with a $\log|\Theta|$ term for general function classes. For the tabular case with finite supported observation and prospective side information, this term scales as $\log|\Theta| = \tilde{O}(M(S^2A + SA|\mathcal{O}|) + M|\mathcal{I}|)$.

4.2. What's Wrong with $\pi^*_{\texttt{blind}}$?

Even if we obtain a sublinear $O(\sqrt{K})$-regret compared to $\pi^*_{\texttt{blind}}$, note that the original goal is to learn the true optimal policy $\pi^* \in \Pi$ which exploits the prospective side information within each trajectory. Therefore, the notion of true regret must be defined in a stronger sense:

$$\text{Regret}(K) = \textstyle\sum_{k=1}^K V_{\theta^*}^{\pi^*} - V_{\theta^*}^{\pi^k}. \quad (4)$$

The overall measure of performance should be on obtaining $\sqrt{K}$-regret with the above stricter definition.

Another issue is, by converting the argument of regret-minimization to sample-complexity, we can obtain ϵ-optimal policy from Algorithm 1 with $\epsilon = O(1/\sqrt{K})$. However, a naive conversion of near-optimal policies in $\Pi_{\texttt{blind}}$ would only guarantee $(|\mathcal{I}|\epsilon)$-optimality for the larger class of policies Π. To see this, suppose O-MLE returns a model θ such that for all $\pi \in \Pi_{\texttt{blind}}$,

$$d_{\text{TV}}\left(\mathbb{P}_\theta^\pi(\iota, \tau), \mathbb{P}_{\theta^*}^\pi(\iota, \tau)\right) \le \epsilon,$$

For the individual ι, however, we can only infer in the worst case that

$$\mathbb{P}_{\theta^*}(\iota) \cdot d_{\text{TV}}\left(\mathbb{P}_\theta^\pi(\tau|\iota), \mathbb{P}_{\theta^*}^\pi(\tau|\iota)\right) \le \min(\mathbb{P}_{\theta^*}(\iota), \epsilon).$$

Thus, when considering a larger policy class $\pi \in \Pi$, a naive analysis would lead to the following upper bound

$$\sum_\iota \mathbb{P}_{\theta^*}(\iota) d_{\text{TV}}\left(\mathbb{P}_{\theta^*}^{\pi(\cdot|\iota)}(\tau|\iota), \mathbb{P}_{\theta^k}^{\pi(\cdot|\iota)}(\tau|\iota)\right) \le \min(1, |\mathcal{I}|\epsilon),$$

since for every ι we use different policy $\pi(\cdot|\iota)$, but a naive analysis would result in a loose bound with multiplicative amplification of the error. Since we consider a large or (almost) continuous observation, the result should not directly depend on $|\mathcal{I}|$, and, instead depend on $\log(|\Theta|)$.

4.3. Hardness of LMDP-Ψ

The first question with prospective side information is whether $\sqrt{K}$-regret is achievable in the stronger sense of Equation (4), i.e., with respect to the stronger comparison policy π^*. Surprisingly (and rather disappointingly), when learning with a larger policy class with the stronger notion of regret, we show that it is impossible to obtain $\sqrt{K}$-regret unless K is larger than $A^{\Omega(M)}$.

Theorem 4.2. *There exists a family of LMDP-Ψs, Θ_{hard}, and a reference model θ_0 with α-prospective side information, such that for any algorithm, the regret of the worst-case instance satisfies with $\alpha < 1/(256\sqrt{M})$,*

$$\inf_{\psi:\texttt{Algs}} \sup_{\theta\in\Theta_{hard}\cup\{\theta_0\}} \sum_{k=1}^{K} V_\theta^{\pi^*} - V_\theta^{\pi^k(\psi)} \gtrsim \min\left(\frac{(A/3)^{(M/4)}}{M\epsilon}, \frac{A}{M\alpha^2\epsilon^2}, \frac{K\epsilon}{M}\right).$$

By optimizing over ϵ, we obtain the following lower bound:

Corollary 4.3. *The regret of any algorithm for the worst-case family of instances satisfy*

$$\inf_{\psi:\texttt{Algs}} \sup_{\theta\in\Theta_{hard}\cup\{\theta_0\}} \sum_{k=1}^{K} V_\theta^{\pi^*} - V_\theta^{\pi^k(\psi)} \gtrsim_P \min\left(A^{\Omega(M)}\sqrt{K}, \left(\frac{A}{\alpha^2}\right)^{1/3} K^{2/3})\right).$$

This lower bound implies the impossibility of designing a learning algorithm with $\mathrm{poly}(M)\sqrt{K}$-regret. Instead, next, we aim to derive an algorithm with an upper bound of $\mathrm{poly}(M)K^{2/3}$ on its regret, i.e., a regret guarantee with no exponential dependence in the number of latent contexts.

4.4. Pure Exploration within $\Pi_{\texttt{blind}}$ is Sufficient

In this section, we present an explore-then-exploit strategy that achieves the optimal $O(K^{2/3})$ regret. When Algorithm 1 (or a reward-free version of it) terminates, the guaranteed inequality is usually on the total variation distance between any model in the confidence set $\theta \in \mathcal{C}^K$ and true model θ^*:

$$\max_{\pi\in\Pi_{\texttt{blind}}} d_{\mathrm{TV}}(\mathbb{P}_{\hat\theta}^\pi, \Pi_{\theta^*}^\pi) \le \epsilon.$$

As discussed earlier, this is not sufficient, and we need a stronger notion of termination criterion, which ensures that all reachable *belief (and the PSR) states* have been sufficiently explored in all models in the remaining confidence. Formally, define the reward bonus for any history at a state-action pair $x := (s, a)$ as

$$\hat\Lambda_t^k(x) = \lambda_0 I + \sum_{j<k} \mathbb{1}\left\{x_t^j = x\right\} \bar b_{\theta^k}(\tau_t^j)\bar b_{\theta^k}(\tau_t^j)^\top,$$
$$\tilde r^k(\tau_t) = \|\bar b_{\theta^k}(\tau_t)\|_{\hat\Lambda_t^k(x)^{-1}},$$

where $\bar b_\theta(\tau_t) = \frac{b_\theta(\tau_t)}{\|b_\theta(\tau_t)\|_1}$ is a normalized PSR of history τ_t in a model θ when a *blind* policy is executed. Our key observation is the following upper bound that relates estimation errors between Π and $\Pi_{\texttt{blind}}$:

$$\max_{\pi\in\Pi} d_{\mathrm{TV}}(\mathbb{P}_{\hat\theta}^\pi, \mathbb{P}_{\theta^*}^\pi) \lesssim_{\mathbb{P}} \max_{\pi\in\Pi_{\texttt{blind}}} \mathbb{E}_{\hat\theta}^\pi\left[\sum_{t=1}^H \tilde r^k(\tau_t)\right]. \quad (5)$$

Algorithm 2 Pure Exploration for LMDP-Ψ

Input: Termination condition $\epsilon_{\texttt{pe}} := \frac{\alpha\epsilon}{10HM^2\sqrt{\lambda_0 M^2/\alpha^2+\beta}}$,
Regularizer $\lambda_0 := \frac{\beta M^2 H^2}{\alpha^2}$

```
Initialize D^0 = ∅, C^0 = Θ
for k = 0...K − 1 do
    # Execute the Worst Blind Policy
    Pick any θ^k ∈ C^k
    π^k = arg max_{π∈Π_blind} Ṽ^π_{θ^k, r̃^k}
    If Ṽ^π_{θ^k, r̃^k} ≤ ε_pe, then break
    Get ι^k and τ^k = (s_1^k, a_1^k, ..., r_H^k) by executing π^k
    # Confidence Set Construction
    D^k ← D^{k−1} ∪ {(τ^k, ι^k, π^k)} and update C^k using (6)
end for
return θ̂ = θ^k
```

A recent result of Huang et al. (2023) (see their Lemma 6) gives an explicit bound on the quantity $\mathbb{E}_{\hat\theta}^\pi\left[\sum_{t=1}^H \tilde r^k(\tau_t)\right]$, instead of bounding the total-variation distance indirectly from the elliptical potential lemma. Therefore, their pure exploration algorithm, *only with policies from a class of blind policies* $\Pi_{\texttt{blind}}$, is sufficient to learn the optimal policy in a larger class of policy Π. We mention that before the result of Huang et al. (2023), direct bound on the cumulative bonus of trajectories did not exist.

Formally, we consider Algorithm 2, where we let $\tau_t := (s_1, a_1, ..., s_t, a_t)$ be a partial trajectory up to time-step t without prospective side information. The expected cumulative bonus at the k^{th} episode in the empirical model is defined as

$$\tilde V_{\theta^k, \tilde r^k}^\pi := \mathbb{E}_{\tau\sim\mathbb{P}_{\theta^k}^\pi}\left[\sum_{t=1}^H \tilde r^k(\tau_t)\right].$$

The confidence set is given based on the likelihood of each model:

$$\mathcal{C}^k := \Big\{\theta\in\Theta \,\Big|\, \textstyle\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k} \log \mathbb{P}_\theta^\pi(\iota,\tau) \ge \max_{\theta'\in\Theta} \sum_{(\iota,\tau,\pi)\in\mathcal{D}^k} \log \mathbb{P}_{\theta'}^\pi(\iota,\tau) - \beta\Big\}. \quad (6)$$

β is pre-defined by the concentration of likelihood value, and is given by $\log(K|\Theta|/\delta)$ as shown in Lemma A.1. Note that from the construction of the confidence set $\mathcal{C}^k$, for all $k \in [K]$, we know that with probability at least $1-\delta$,

$$-\sum_{(\tau,\pi)\in\mathcal{D}^k} \log\left(\frac{\mathbb{P}_{\theta^k}^\pi(\iota,\tau)}{\mathbb{P}_{\theta^*}^\pi(\iota,\tau)}\right) \le 2\beta.$$

Thus, we may simply choose the maximum likelihood estimator (MLE). We obtain the following guarantee:

Theorem 4.4. *Let $\epsilon_{\texttt{pe}}, \lambda_0$ as defined in the input in Algorithm 2. Then, with probability at least $1-\delta$, Algorithm 2*

returns a model $\hat{\theta}$ after at most K episodes where

$$K = O\left(\frac{M^8 H^4 SA \cdot \log(K|\Theta|/\delta)\log(K)}{\alpha^6 \epsilon^2}\right), \quad (7)$$

Furthermore, the optimal policy $\pi^*_{\hat{\theta}} \in \Pi$ for the returned model $\hat{\theta}$ is an ϵ-optimal policy for θ^* with probability at least $1-\delta$, i.e., $\left|V^{\pi^*}_{\theta^*} - V^{\pi^*_{\hat{\theta}}}_{\theta^*}\right| \le \epsilon$.

Finally, the sample complexity guarantee can naturally be converted into a regret guarantee by a standard explore-then-exploit approach. That is, by playing $\epsilon^{-2} = O(K^{2/3})$ to obtain an ϵ-optimal policy and exploit the learned policy for the remaining episode. For regret minimization, we get:

$$\sum_{k=1}^{K} V^{\pi^*}_{\theta^*} - V^{\pi^k}_{\theta^*} \lesssim \left(\frac{M^8 H^4 SA \cdot \log(|\Theta|/\delta)}{\alpha^6}\right)^{1/3} K^{2/3},$$

regret bound up to logarithmic factors for K episodes.

5. Analysis

In this section, we provide the upper and lower bounds proofs and intuition.

5.1. Upper Bound

Here, we provide the overview of analyzing Algorithm 2. The main step is to establish the inequality of equation (5). We adopt the idea from Huang et al. (2023) of separating the concentration argument (for bounding the sum of TV distances) and the elliptical potential argument. In addition to the notation defined in equation (2), we let

$$b(\tau_t) := \Pi_{h=1}^{t-1} B(o_h, s_{h+1}|s_h, a_h) b_0,$$
$$\pi(\iota, \tau_t) := \Pi_{h=1}^{t} \pi(a_h|\iota, s_1, ..., s_h),$$
$$\pi(\omega_t|\iota, \tau_t) := \Pi_{h=t+1}^{H} \pi(a_h|\iota, s_1, ..., s_h).$$

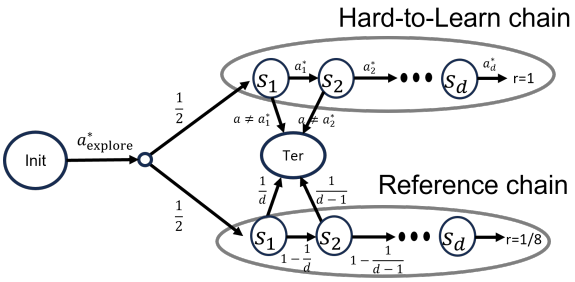
Our crucial observation on exploiting the prospective weakly revealing side information is the following conditional, on the value of ι, well conditioning of the LMDP-Ψ system:

Lemma 5.1. *Fix any prospective side information $\iota \in \mathcal{I}$. For all $x_t = (s_t, a_t) \in \mathcal{S} \times \mathcal{A}$, $t \in [H]$, and π that is independent of the history before time-step t, we have*

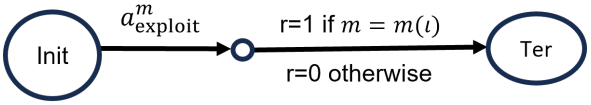
$$\max_{b:\|b\|_1=1} \max_{\pi} \sum_{\omega_t} \pi(\omega_t)|\psi(\omega_t, \iota|x_t)^\top b| \le \frac{M}{\alpha} \max_{m\in[M]} \mathbb{P}(\iota|m).$$

On the other hand, following the standard algebra to bound the total variation distance, we can bound $d_{\mathrm{TV}}(\mathbb{P}^{\pi}_{\theta^*}, \mathbb{P}^{\pi}_{\theta})$ as for all θ as follows:

$$d_{\mathrm{TV}}(\mathbb{P}^{\pi}_{\theta^*}, \mathbb{P}^{\pi}_{\theta}) \le \sum_{t=1}^{H} \sum_{\iota,\tau_t} \pi(\tau_t|\iota) \sum_{\omega_t} |f(\omega_t, \iota) b_\theta(\tau_t)|,$$



Case I: ι_{hard} is hardly informative



Case II: $\iota \neq \iota_{\text{hard}}$ nearly specifies the context

Figure 1: Optimal behaviors in the family of hard instances. The numbers on the arrow mean the probability of transitions under the optimal policy. Actions on the arrow mean transition happens when the action satisfying the condition is taken.

where $f(\omega_t, \iota) := \pi(\omega_t|\tau_t, \iota) \cdot \psi_{\theta^*}(\omega_{t+1}, \iota|x_{t+1})^\top (B_{\theta^*}(y_t|x_t) - B_\theta(y_t|x_t))$ is the term involving the operator difference between two models. The inner summation over the partial future trajectory ω_t can be split into the multiplication of the concentration error *conditioned* on ι:

$$\|\textstyle\sum_{\omega_t} f(\omega_t, \iota)\|_{\hat{\Lambda}_t(x_t)}$$

and the cumulative sum of trajectory bonuses when the prospective side information ι is ignored:

$$\|\bar{b}_\theta(\tau_t)\|_{\hat{\Lambda}_t(x_t)^{-1}}.$$

For the concentration error in PSRs, we can apply the conditional concentration of total-variation distances for likelihood estimators (see Appendix A.3) and Lemma 5.1. For the cumulative bonuses, we use the termination condition of Algorithm 2. Combining the two, we can prove Theorem 4.4. See Appendix B for the complete proofs.

5.2. Lower Bound

Next, we describe the lower bound construction and supply intuition for this result. Consider the following scenario (see Figure 1 for the class of LMDP-Ψs): suppose that for a non-negligible portion of episodes, the prospective side information does supply any information on the latent context. That is, given the prospective side information ι_{hard} the posterior probability over the latent contexts is uniform, *i.e.,* $\mathbb{P}(m|\iota_{\text{hard}}) = 1/M$, and ι_{hard} happens with constant probability, e.g., $1/4$. With ι_{hard} alone, however, learning the optimal action sequence $a^*_{1:d}$ (optimal policy)

may suffer from an exponential lower bound $A^{\Omega(M)}$ since ι_{hard} supplies no information on the latent context. At the same time, playing any sub-optimal action sequence incurs an $\Omega(\epsilon)$ regret where ϵ is the required target accuracy.

On the other hand, any other prospective side information $\iota \neq \iota_{\text{hard}}$, provides a strong signal of one environment that is the most likely, *i.e.,* $\mathbb{P}(m^*(\iota)|\iota) \geq 1/2$. Further, suppose that there is a unique exploiting action for each context that always gives a high reward, and playing any other action incurs $O(1)$-regret. For this environment, the regret of any algorithm is proportional to how many times the sub-optimal action is played when $\iota \neq \iota_{\text{hard}}$.

However, it is still essential to learn the optimal sequence of actions $a^*_{1:d}$ in order to behave optimally under ι_{hard}. Therefore, to avoid the exponential lower bound, we should be aided by good prospective side information $\iota \neq \iota_{\text{hard}}$ despite the strong signal of the underlying model. We can construct internal dynamics such that we need to explore the two chains for at least $\Omega(A/\alpha^2\epsilon^2)$ episodes to identify the optimal action sequence $a^*_{1:d}$ when $\iota \neq \iota_{\text{hard}}$. Combining these arguments, the regret lower bound should be at least $\min\left(\frac{A^{\Omega}(M)}{\epsilon}, \frac{A}{\alpha^2\epsilon^2}, K\epsilon\right)$, yielding Theorem 4.2.

To obtain the multiplicative dependence on α and ϵ, the actual construction of the hard instance family is slightly more complicated. We assume that M is sufficiently large and a multiple of 4, and let $d = M/4$. We also assume that $\alpha \ll 1$ is a sufficiently small constant. We let the time step start from $t = 0$ at the initial state s_{init}. Next, we describe the construction of the hard LMDP-Ψ class:

State Space. There are four categories of states. The initial state s_{init}, absorbing state s_{ter} (which means essentially an episode is terminated), a chain of states constructing a hard-to-learn system $s^{\text{hard}}_{1:d}$, and another chain of states constructing a reference system $s^{\text{ref}}_{1:d}$.

Action Space. The set of actions at the initial time step consists of a set of candidate exploring actions $\mathcal{A}_{\text{explore}}$ and exploiting actions $\mathcal{A}_{\text{exploit}} := \{a^m_{\text{exploit}}\}^M_{m=M/2+1}$ that control the dynamics at the initial state. The action set at time steps $1, \cdots, d$, denoted by $\mathcal{A}_{\text{control}}$, controls the dynamics in hard-to-learn and reference chains of the system. At the initial time step, only one action of $\mathcal{A}_{\text{explore}}$ is a true exploring action a^*_{explore}. At time steps $1, \cdots, d$ only one action sequence $a^*_{1:d} \in \mathcal{A}^{\bigotimes d}_{\text{control}}$ is the optimal sequence.

Latent Environments and Initial Dynamics. There are three groups of MDPs: $\mathcal{G}_{\texttt{learn}}$, $\mathcal{G}_{\texttt{ref}}$, and $\mathcal{G}_{\texttt{obs}}$. All MDPs always start from the same starting state s_{init}.

$\mathcal{G}_{\texttt{learn}}$ consists of $(M/4)$ MDPs, $\mathcal{M}_1, ..., \mathcal{M}_{M/4}$, which essentially form the hard to learn example from Kwon et al. (2021) when no prospective side information is provided. In any of these environments, in the beginning, when the 'true' explore action a^*_{explore} is played, it transitions to the starting of hard-instance chain s^{hard}_1 with some small probability.

Similarly, $\mathcal{G}_{\texttt{ref}}$ consists of another $(M/4)$ MDPs, $\mathcal{M}_{M/4+1}, ..., \mathcal{M}_{M/2}$, and the purpose of $\mathcal{G}_{\texttt{ref}}$ is to confuse the learning the optimal action sequence in the hard-to-learn chain, as we make the prospective side information hard to distinguish whether an MDP belongs to $\mathcal{G}_{\texttt{learn}}$ or $\mathcal{G}_{\texttt{ref}}$. More precisely, under ι_{hard}, it is hard to identify which one is the hard-to-learn or reference chain, and thus it is hard to identify a^*_{explore}. This is crucial to build a multiplicative lower bound on α and ϵ.

The rest of $(M/2)$ MDPs, indexed by $\mathcal{M}_{M/2+1}, ..., \mathcal{M}_M$, belong to the almost observable group $\mathcal{G}_{\texttt{obs}}$. In each environment of this group $\mathcal{M}_m \in \mathcal{G}_{\texttt{obs}}$ where $m = M/2 + 1, ..., M$, executing a^m_{exploit} at the initial time step step results with a reward 1, and gets 0 otherwise.

Dynamics of Two Chains. In both hard and reference chains, at any states in $s^{\text{hard}}_{1:d}$ and $s^{\text{ref}}_{1:d}$, all actions $a \notin \mathcal{A}_{\text{control}}$ invoke transitions to s_{ter} with 0 rewards.

In the reference chain, in all environments in $\mathcal{G}_{\texttt{learn}} \cup \mathcal{G}_{\texttt{ref}}$, for all actions $a \in \mathcal{A}_{\text{control}}$, s^{ref}_t transitions to s^{ref}_{t+1} with probability $\left(1 - \frac{1}{d+1-t}\right)$ and transitions to s^{ter} otherwise when $t < d$. When the chain transitions to s_{ter}, we receive a reward sampled from $\text{Ber}(1/8)$.

In the hard-to-learn chain, for all environments in $\mathcal{G}_{\texttt{ref}}$, the system dynamic is identical to the reference chain. The environments in $\mathcal{G}_{\texttt{learn}}$ are set to be the hard family instances of MDPs from Kwon et al. (2021) (while setting $d = M/4$), also depicted in Figure 1, Case I:

1. At each time, MDPs in $\mathcal{G}_{\texttt{learn}}$ transitions from one state in the chain to the next state or to s_{ter} depending on the played action. When an agent transitions to s_{ter} it receives a reward drawn from $\text{Ber}(1/8)$.

2. At all time steps besides at the last one, the agent receives a reward of 0, when taking an action that does not take it to s_{ter}. At the last time step, if the agent did not move to s_{ter} and upon taking the action a^*_d it recives a reward of 1. Hence, the essence of this construction is to identify the optimal action sequence $a^*_{1:d}$ which guarantees a reward 1 from $\mathcal{M}_1$ at the end of the chain s^{hard}_d. Playing any sub-optimal action sequence generates the distribution of observations indistinguishable from the reference chain.

We complete the construction in Appendix C.

Prospective Side Information. The prospective side information either is a strong prior of one of the MDPs in $\mathcal{G}_{\texttt{obs}}$, or uninformative in which case $\iota = \iota_{\text{hard}}$. Our construction ensures that when observing ι_{hard}, all MDPs in $\mathcal{G}_{\text{learn}}$ and $\mathcal{G}_{\text{ref}}$ have equal conditional probability, *i.e.,* $\mathbb{P}(m|\iota) = 2/M$ for all $m \in [M/2]$, whereas for other values of prospective information $\iota \neq \iota_{\text{hard}}$, there is one MDP from $\mathcal{G}_{\text{obs}}$ whose prior probability is greater than $1/2$, and priors over $\mathcal{G}_{\text{learn}} \cup \mathcal{G}_{\text{ref}}$ are nearly equally distributed but perturbed by a small parameter α, *i.e.,* $\mathbb{P}(m_{\text{obs}}|\iota) \geq 1/2$ for some $m_{\text{obs}} \in [M/2+1, M]$, and $\mathbb{P}(m|\iota) = O(1/M) + O(\alpha)$ for all $m \in [M/2]$.

Hard Instances. The family of hard instances Θ_{hard} that consists the set of hard-to-learn LMDP-Ψs is described as follows. All instances in the hard instance family share the same state space, action space and prospective side-information. The family of hard-to-learn LMDP-Ψs differ in their transition dynamics. Each LMDP-Ψ in Θ_{hard} differs by its transition dynamics. The transition dynamics of each element of Θ_{hard} is determined by one of the possible sequences $a^*_{1:d} \in \mathcal{A}_{\text{control}}^{\otimes d}$ that represents the optimal action sequence, and by the 'true' exploring actions $a^*_{\text{explore}} \in \mathcal{A}_{\text{explore}}$.

Reference Model. We denote θ_0 as the reference model whose hard-to-learn chain is no different from the reference chain in all individual MDPs. In the reference model, at s_{init}, all MDPs in $\mathcal{G}_{\texttt{ref}}$ transitions to s_1^{hard} and those in $\mathcal{G}_{\texttt{learn}}$ transitions to s_1^{ref} deterministically when any action in $\mathcal{A}_{\text{explore}}$ is played. All other parts are constructed with the same dynamics as in Θ_{hard}.

Proof Overview. With the above construction, the following lemmas play key roles in proving the lower bound:

Lemma 5.2. *Let ψ be any exploration strategy for LMDP-Ψ. Consider any hard instance $\theta \in \Theta_{hard}$ and the reference model θ_0. Let $N^{explore}_{\psi,\iota,a_{1:d}}(K)$ be the number of times that explored the chain systems with the test $t_\iota(a_{1:d}) := \{\iota, a^*_{explore}, a_{1:d}\}$, i.e., with the true exploration action and any sequence $a_{1:d} \in \mathcal{A}^{\otimes d}$ given prospective side information ι. Then,*

$$\sum_{\iota, a_{1:d}} \mathbb{E}_{\theta_0}\left[N^{explore}_{\psi,\iota,a_{1:d}}(K)\right] \cdot KL\left(\mathbb{P}_{\theta_0}(\cdot|t_\iota(a_{1:d})), \mathbb{P}_\theta(\cdot|t_\iota(a_{1:d}))\right) = KL\left(\mathbb{P}^\psi_{\theta_0}(\tau^{1:K}), \mathbb{P}^\psi_\theta(\tau^{1:K})\right), \tag{8}$$

where $\mathbb{P}^\psi(\tau^{1:K})$ is a distribution of K trajectories obtained with the exploration strategy ψ.

The main reason for the equality (8) is that whenever $a \neq a^*_{\text{explore}}$ is played regardless of the prospective side information, the two models θ and θ_0 generate observations from the same distribution. Then the key lemma is on the bounds for the conditional KL-divergence:

Lemma 5.3. *For all non optimal action sequences $a_{1:d} \neq a^*_{1:d}$, the following holds:*

$$KL\left(\mathbb{P}_{\theta_0}(\cdot|\iota_{hard}, a^*_{explore}, a_{1:d}), \mathbb{P}_\theta(\cdot|\iota_{hard}, a^*_{explore}, a_{1:d})\right) = 0,$$

and for all $\iota \in \mathcal{I}$,

$$KL\left(\mathbb{P}_{\theta_0}(\cdot|\iota, a^*_{explore}, a^*_{1:d}), \mathbb{P}_\theta(\cdot|\iota, a^*_{explore}, a^*_{1:d})\right) \lesssim \epsilon^2.$$

Furthermore, for all $\iota \neq \iota_{hard}$ and $a_{1:d} \neq a^*_{1:d}$:

$$KL\left(\mathbb{P}_{\theta_0}(\cdot|\iota, a^*_{explore}, a_{1:d}), \mathbb{P}_\theta(\cdot|\iota, a^*_{explore}, a_{1:d})\right) \lesssim (\alpha\epsilon)^2.$$

Therefore, we can bound the KL-divergence between the total trajectory distributions of the two models as

$$\begin{aligned}&\text{KL}\left(\mathbb{P}^\psi_{\theta_0}(\tau^{1:K}), \mathbb{P}^\psi_\theta(\tau^{1:K})\right)\\ &\leq \mathbb{E}_{\theta_0}[N^{\text{explore}}_{\psi,\iota_{\text{hard}},a^*_{1:d}}]\epsilon^2 + \sum_{\iota\neq\iota_{\text{hard}},a_{1:d}} \mathbb{E}_{\theta_0}[N^{\text{explore}}_{\psi,\iota,a_{1:d}}](\alpha\epsilon)^2,\end{aligned}$$

which translates to the impossibility of distinguishing the two with a probability more than $2/3$ unless either

$$\begin{aligned}&\mathbb{E}_{\theta_0}[N^{\text{explore}}_{\psi,\iota_{\text{hard}},a^*_{1:d}}(K)] \gtrsim \frac{1}{\epsilon^2},\\ &\text{or} \sum_{\iota\neq\iota_{\text{hard}},a_{1:d}} \mathbb{E}_{\theta_0}[N^{\text{explore}}_{\psi,\iota,a_{1:d}}(K)] \gtrsim \frac{1}{\alpha^2\epsilon^2}.\end{aligned} \tag{9}$$

Finally, note that playing sub-optimal actions with $\iota \neq \iota_{\text{hard}}$ incurs at least $1/8$-regret, playing sub-optimal action sequence $a_{1:d} \neq a^*_{1:d}$ incurs at least $O(\epsilon/M)$-regret, and playing the optimal sequence $a^*_{1:d}$ at least $O(1/\epsilon^2)$ times would take (A^d/ϵ^2) episodes in the worst case. The remaining steps are to formally state the ideas (see Appendix C).

6. Conclusion

In this work, we introduced the LMDP-Ψ setting, when a prospective and weakly revealing information on the latent context is given to an agent. We showed that LMDP-Ψ does not belong to the weakly revealing POMDP class, as our results highlight its fundamental different characteristic: for an LMDP-Ψ, differently than a weakly revealing POMDP, an $O(\sqrt{K})$ worst-case upper bound is not achievable without suffering exponential dependence in the problem parameters. We complemented this negative result with a positive one: an $\Omega(K^{2/3})$ lower bound and a matching upper bound that depends *polynomially* on problem parameters.

From a broader perspective, our results highlight a key deficiency of a ubiquitous assumption made in POMDP modeling, namely, *the independence of observation* between consecutive time steps, when conditioning on the latent state.

This assumption is violated in the presence of prospective information and, specifically, in the LMDP-Ψ setting. We believe that studying the learnability of more general POMDP settings with prospective side information, or non-trivial correlation between observations serves as a fruitful ground for future work. Further, scaling the methods for practical settings, while building on a solid theoretical foundation, is a valuable and open research direction.

Impact statement. This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgement

This research was partially funded by AFOSR/AFRL grant no. FA9550-18-1-0166, NSF Grants 2019844 and 2112471, and Israel Science Foundation Grant No. 2199/20.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

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.

Brunskill, E. and Li, L. Sample complexity of multi-task reinforcement learning. In *Uncertainty in Artificial Intelligence*, pp. 122. Citeseer, 2013.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Chadès, I., Carwardine, J., Martin, T., Nicol, S., Sabbadin, R., and Buffet, O. MOMDPs: a solution for modelling adaptive management problems. In *Twenty-Sixth AAAI Conference on Artificial Intelligence (AAAI-12)*, 2012.

Chatterjee, K., Chmelík, M., Karkhanis, D., Novotnỳ, P., and Royer, A. Multiple-environment markov decision processes: Efficient analysis and applications. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 30, pp. 48–56, 2020.

Chen, F., Bai, Y., and Mei, S. Partially observable rl with b-stability: Unified structural condition and sharp sample-efficient algorithms. In *The Eleventh International Conference on Learning Representations*, 2022.

Chen, F., Wang, H., Xiong, C., Mei, S., and Bai, Y. Lower bounds for learning in revealing pomdps. *arXiv preprint arXiv:2302.01333*, 2023.

Garivier, A., Ménard, P., and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

Golowich, N., Moitra, A., and Rohatgi, D. Learning in observable pomdps, without computationally intractable oracles. *Advances in Neural Information Processing Systems*, 35:1458–1473, 2022.

Hallak, A., Di Castro, D., and Mannor, S. Contextual markov decision processes. *arXiv preprint arXiv:1502.02259*, 2015.

Huang, R., Liang, Y., and Yang, J. Provably efficient ucb-type algorithms for learning predictive state representations. *arXiv preprint arXiv:2307.00405*, 2023.

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.

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, 2021.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. Tractable optimality in episodic latent mabs. *Advances in Neural Information Processing Systems*, 35:23634–23645, 2022.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. Reward-mixing mdps with few latent contexts are learnable. In *International Conference on Machine Learning*, pp. 18057–18082. PMLR, 2023.

Lee, J., Agarwal, A., Dann, C., and Zhang, T. Learning in pomdps is sample-efficient with hindsight observability. In *International Conference on Machine Learning*, pp. 18733–18773. PMLR, 2023.

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

Liu, Q., Netrapalli, P., Szepesvari, C., and Jin, C. Optimistic mle: A generic model-based algorithm for partially observable sequential decision making. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 363–376, 2023.

Liu, Y., Guo, Z., and Brunskill, E. PAC continuous state online multitask reinforcement learning with identification. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*, pp. 438–446, 2016.

Modi, A., Jiang, N., Singh, S., and Tewari, A. Markov decision processes with continuous side information. In *Algorithmic Learning Theory*, pp. 597–618. PMLR, 2018.

Paley, R. and Zygmund, A. On some series of functions,(1). In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 26, pp. 337–357. Cambridge University Press, 1930.

Steimle, L. N., Kaufman, D. L., and Denton, B. T. Multi-model markov decision processes. *Optimization Online URL http://www. optimization-online. org/DB_FILE/2018/01/6434. pdf*, 2018.

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.

Taylor, M. E. and Stone, P. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(7), 2009.

Uehara, M., Sekhari, A., Lee, J. D., Kallus, N., and Sun, W. Provably efficient reinforcement learning in partially observable dynamical systems. *arXiv preprint arXiv:2206.12020*, 2022.

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

Supplementary Materials for "Prospective Side Information for Latent MDPs"

A. Auxiliary Lemmas

Lemma A.1 (General MLE, Liu et al. (2022)). *With probability* $1-\delta$ *for any* $\delta > 0$, *for all* $k \in [K]$, $t \in [H]$ *and for any* $\theta \in \Theta$,

$$\sum_{(\iota,\tau_t,\pi)\in\mathcal{D}^k} \log(\mathbb{P}_\theta^\pi(\iota,\tau_t)) - 3\log(K|\Theta|/\delta) \le \sum_{(\iota,\tau_t,\pi)\in\mathcal{D}^k} \log(\mathbb{P}_{\theta^*}^\pi(\iota,\tau_t)). \tag{10}$$

This is by now a standard MLE technique for constructing confidence sets in RL (Agarwal et al., 2020).

Proof. The proof follows a Chernoff bound type of technique:

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\sum_{(\iota,\tau_t,\pi)\in\mathcal{D}^k} \log\left(\frac{\mathbb{P}_\theta^\pi(\iota,\tau_t)}{\mathbb{P}_{\theta^*}^\pi(\iota,\tau_t)}\right) \ge \mathbb{E}_{\theta^*}\left[\sum_{(\iota,\tau_t,\pi)\in\mathcal{D}^k} \log\left(\frac{\mathbb{P}_\theta^\pi(\iota,\tau_t)}{\mathbb{P}_{\theta^*}^\pi(\iota,\tau_t)}\right)\right] + \beta\right) \\
&\le \mathbb{P}_{\theta^*}\left(\exp\left(\sum_{(\iota,\tau_t,\pi)\in\mathcal{D}^k} \log\left(\frac{\mathbb{P}_\theta^\pi(\iota,\tau_t)}{\mathbb{P}_{\theta^*}^\pi(\iota,\tau_t)}\right)\right) \ge \exp(\beta)\right) \\
&\le \mathbb{E}_{\theta^*}\left[\exp\left(\sum_{(\iota,\tau_t,\pi)\in\mathcal{D}^k} \log\left(\frac{\mathbb{P}_\theta^\pi(\iota,\tau_t)}{\mathbb{P}_{\theta^*}^\pi(\iota,\tau_t)}\right)\right)\right]\exp(-\beta).
\end{aligned}$$

The last inequality is by Markov's inequality. Note that random variables are (ι,τ_t,π) in the trajectory dataset $\mathcal{D}^k$, and

$$\mathbb{E}_{\theta^*}\left[\sum_{(\tau,\pi)\in\mathcal{D}^k} \log\left(\frac{\mathbb{P}_\theta^\pi(\iota,\tau_t)}{\mathbb{P}_{\theta^*}^\pi(\iota,\tau_t)}\right)\right] = -\mathrm{KL}(\mathbb{P}_{\theta^*}(\mathcal{D}^k)||\mathbb{P}_\theta(\mathcal{D}^k)) \le 0.$$

Then,

$$\mathbb{E}_{\theta^*}\left[\exp\left(\sum_{(\iota,\tau_t,\pi)\in\mathcal{D}^k} \log\left(\frac{\mathbb{P}_\theta^\pi(\iota,\tau_t)}{\mathbb{P}_{\theta^*}^\pi(\iota,\tau_t)}\right)\right)\right] = \mathbb{E}_{\theta^*}\left[\Pi_{(\iota,\tau_t,\pi)\in\mathcal{D}^k}\frac{\mathbb{P}_\theta^\pi(\iota,\tau_t)}{\mathbb{P}_{\theta^*}^\pi(\iota,\tau_t)}\right] = \sum_{\mathcal{D}^k}\mathbb{P}_\theta(\mathcal{D}^k) = 1.$$

Combining the above, taking a union bound over $k \in [K]$ and $\theta \in \Theta$, letting $\beta = \log(K|\Theta|/\delta)$, with probability $1-\delta$, the inequality in equality (10) holds. □

Lemma A.2. *With probability* $1-\delta$, *for all* $k \in [K]$, $t \in [H]$ *and* $\theta \in \Theta$, *we have*

$$\begin{aligned}
\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k} d_{TV}^2(\mathbb{P}_\theta^\pi(\iota,\tau), \mathbb{P}_{\theta^*}^\pi(\iota,\tau)) &\lesssim \sum_{(\iota,\tau,\pi)\in\mathcal{D}^k} \log\left(\frac{\mathbb{P}_{\theta^*}^\pi(\iota,\tau)}{\mathbb{P}_\theta^\pi(\iota,\tau)}\right) + \beta, \\
\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k} d_H^2(\mathbb{P}_\theta^\pi(\iota,\tau), \mathbb{P}_{\theta^*}^\pi(\iota,\tau)) &\lesssim \sum_{(\iota,\tau,\pi)\in\mathcal{D}^k} \log\left(\frac{\mathbb{P}_{\theta^*}^\pi(\iota,\tau)}{\mathbb{P}_\theta^\pi(\iota,\tau)}\right) + \beta.
\end{aligned}$$

Proof. By the TV-distance and Hellinger distance relation, for any ι, τ, π and $t \in [H]$,

$$\begin{aligned}
d_{\mathrm{TV}}^2(\mathbb{P}_\theta^\pi(\iota,\tau), \mathbb{P}_{\theta^*}^\pi(\iota,\tau)) &\le 2d_{\mathrm{H}}^2(\mathbb{P}_\theta^\pi(\iota,\tau), \mathbb{P}_{\theta^*}^\pi(\iota,\tau)) \\
&= 2\left(1 - \mathbb{E}_{\iota,\tau\sim\mathbb{P}_{\theta^*}^\pi}\left[\sqrt{\frac{\mathbb{P}_\theta^\pi(\iota,\tau)}{\mathbb{P}_{\theta^*}^\pi(\iota,\tau)}}\right]\right)
\end{aligned}$$

$$\leq -2\log\left(\mathbb{E}_{\iota,\tau\sim\mathbb{P}^{\pi}_{\theta^*}}\left[\sqrt{\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}}\right]\right).$$

To bound the summation over samples, we start from

$$\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k} d^2_{\mathrm{TV}}\left(\mathbb{P}^{\pi}_{\theta}(\iota,\tau),\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)\right) \leq -2\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k}\log\left(\mathbb{E}_{\iota,\tau\sim\mathbb{P}^{\pi}_{\theta^*}}\left[\sqrt{\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}}\right]\right).$$

On the other hand, by the Chernoff bound,

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\left(\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k}\log\left(\sqrt{\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}}\right) \geq \sum_{(\iota,\tau,\pi)\in\mathcal{D}^k}\log\mathbb{E}_{\iota,\tau\sim\mathbb{P}^{\pi}_{\theta^*}}\left[\sqrt{\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}}\right]+\beta\right)\\
&\leq \mathbb{E}_{\theta^*}\left[\frac{\exp\left(\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k}\log\left(\sqrt{\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}}\right)\right)}{\exp\left(\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k}\log\mathbb{E}_{\iota,\tau\sim\mathbb{P}^{\pi}_{\theta^*}}\left[\sqrt{\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}}\right]\right)}\right]\exp(-\beta)\\
&= \mathbb{E}_{\theta^*}\left[\frac{\Pi_{(\iota,\tau,\pi)\in\mathcal{D}^k}\sqrt{\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}}}{\Pi_{(\iota,\tau,\pi)\in\mathcal{D}^k}\mathbb{E}_{\iota,\tau\sim\mathbb{P}^{\pi}_{\theta^*}}\left[\sqrt{\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}}\right]}\right]\exp(-\beta)\\
&= \mathbb{E}_{\theta^*}\left[\frac{\Pi_{(\iota,\tau,\pi)\in\mathcal{D}^{k-1}}\sqrt{\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}}\cdot\mathbb{E}_{\iota,\tau^k\sim\mathbb{P}^{\pi^k}_{\theta^*}}\left[\sqrt{\frac{\mathbb{P}^{\pi^k}_{\theta}(\iota,\tau^k)}{\mathbb{P}^{\pi^k}_{\theta^*}(\iota,\tau^k)}}\middle|\pi^k,\mathcal{D}^{k-1}\right]}{\Pi_{(\iota,\tau,\pi)\in\mathcal{D}^k}\mathbb{E}_{\iota,\tau\sim\mathbb{P}^{\pi}_{\theta^*}}\left[\sqrt{\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}}\right]}\right]\exp(-\beta)\\
&= \mathbb{E}_{\theta^*}\left[\frac{\Pi_{(\iota,\tau,\pi)\in\mathcal{D}^{k-1}}\sqrt{\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}}}{\Pi_{(\iota,\tau,\pi)\in\mathcal{D}^{k-1}}\mathbb{E}_{\iota,\tau\sim\mathbb{P}^{\pi}_{\theta^*}}\left[\sqrt{\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}}\right]}\right]\exp(-\beta)=\ldots=\exp(-\beta),
\end{aligned}$$

where in the last line, we used the tower property of expectation. Thus, again by setting $\beta = O\left(\log(KH|\Theta|/\delta)\right)$, with probability at least $1-\delta$, we have

$$\begin{aligned}
\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k} d^2_{\mathrm{TV}}(\mathbb{P}^{\pi}_{\theta}(\iota,\tau),\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)) &\lesssim -\frac{1}{2}\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k}\log\left(\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}\right)+\beta\\
&= -\frac{1}{2}\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k}\log\left(\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}\right)+\frac{1}{2}\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k}\log\left(\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}\right)+\beta.
\end{aligned}$$

We can apply Lemma A.1, and finally have

$$\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k} d^2_{\mathrm{TV}}(\mathbb{P}^{\pi}_{\theta}(\iota,\tau),\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)) \lesssim -\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k}\log\left(\frac{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}\right)+\beta.$$

□

Most of the following lemmas can also be found in (Huang et al., 2023) as we adopt their proof strategy. We state and prove them for the completeness. The following is the concentration lemma for the empirical *conditional* probability, which Importantly, this property still holds regardless of causal relationships inside each trajectory sample:

Lemma A.3. *With probability* $1-\delta$*, for all* $k\in[K]$*,* $t\in[H]$*,* $\theta\in\Theta$*, we have*

$$\begin{aligned}
\sum_{(\iota,\tau_t,\omega_t,\pi)\in\mathcal{D}^k} d^2_{TV}(\mathbb{P}^{\pi}_{\theta}(\iota,\omega_t|\tau_t),\mathbb{P}^{\pi}_{\theta^*}(\iota,\omega_t|\tau_t)) &\lesssim \sum_{(\iota,\tau,\pi)\in\mathcal{D}^k}\log\left(\frac{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}\right)+\beta,\\
\sum_{(\iota,\tau_t,\omega_t,\pi)\in\mathcal{D}^k} d^2_{H}(\mathbb{P}^{\pi}_{\theta}(\iota,\omega_t|\tau_t),\mathbb{P}^{\pi}_{\theta^*}(\iota,\omega_t|\tau_t)) &\lesssim \sum_{(\iota,\tau,\pi)\in\mathcal{D}^k}\log\left(\frac{\mathbb{P}^{\pi}_{\theta^*}(\iota,\tau)}{\mathbb{P}^{\pi}_{\theta}(\iota,\tau)}\right)+\beta.
\end{aligned}$$

Proof. The proof is almost identical except that we now start from

$$\sum_{(\tau,\pi)\in\mathcal{D}^k} d_{\mathrm{TV}}^2\left(\mathbb{P}_\theta^\pi(\iota,\omega_t|\tau_t),\mathbb{P}_{\theta^*}^\pi(\iota,\omega_t|\tau_t)\right) \le -2\sum_{(\tau,\pi)\in\mathcal{D}^k}\log\left(\mathbb{E}_{(\iota,\omega_t)\sim\mathbb{P}_{\theta^*}^\pi(\cdot|\tau_t)}\left[\sqrt{\frac{\mathbb{P}_\theta^\pi(\iota,\omega_t|\tau_t)}{\mathbb{P}_{\theta^*}^\pi(\iota,\omega_t|\tau_t)}}\right]\right).$$

and use the tower property of expectation conditioned on τ_t^k. Thus, again by setting $\beta = O\left(\log(KH|\Theta|/\delta)\right)$, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sum_{(\tau,\pi)\in\mathcal{D}^k} d_{\mathrm{TV}}^2\left(\mathbb{P}_\theta^\pi(\iota,\omega_t|\tau_t),\mathbb{P}_{\theta^*}^\pi(\iota,\omega_t|\tau_t)\right) \lesssim -\frac{1}{2}\sum_{(\tau,\pi)\in\mathcal{D}^k}\log\left(\frac{\mathbb{P}_\theta^\pi(\iota,\omega_t|\tau_t)}{\mathbb{P}_{\theta^*}^\pi(\iota,\omega_t|\tau_t)}\right)+\beta\\
&=-\frac{1}{2}\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k}\log\left(\frac{\mathbb{P}_\theta^\pi(\iota,\tau)}{\mathbb{P}_{\theta^*}^\pi(\iota,\tau)}\right)+\frac{1}{2}\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k}\log\left(\frac{\mathbb{P}_\theta^\pi(\tau_t)}{\mathbb{P}_{\theta^*}^\pi(\tau_t)}\right)+\beta.
\end{aligned}$$

Finally, we apply Lemma A.1, and have

$$\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k} d_{\mathrm{TV}}^2\left(\mathbb{P}_\theta^\pi(\iota,\omega_t|\tau_t),\mathbb{P}_{\theta^*}^\pi(\iota,\omega_t|\tau_t)\right) \lesssim -\sum_{(\iota,\tau,\pi)\in\mathcal{D}^k}\log\left(\frac{\mathbb{P}_\theta^\pi(\iota,\tau)}{\mathbb{P}_{\theta^*}^\pi(\iota,\tau)}\right)+\beta.$$

□

Lemma A.4. *For arbitrary probability distribution P, Q over joint distributions (τ,ω),*

$$\mathbb{E}_{\tau\sim P}[d_H^2(P(\omega|\tau),Q(\omega|\tau))] \le 4d_H^2(P(\omega,\tau),Q(\omega,\tau)).$$

Proof. We prove this statement by explicitly bounding the Hellinger distance.

$$\begin{aligned}
&\int\left(\int\left(\sqrt{P(\omega|\tau)}-\sqrt{Q(\omega|\tau)}\right)^2 d\omega\right)P(\tau)d\tau\\
&\le 2\int\int\left(\sqrt{P(\omega,\tau)}-\sqrt{Q(\tau)Q(\omega|\tau)}\right)^2 d\omega d\tau+2\int\int\left(\sqrt{P(\tau)Q(\omega|\tau)}-\sqrt{Q(\tau)Q(\omega|\tau)}\right)^2 d\omega d\tau\\
&=2d_{\mathrm{H}}^2(P(\omega,\tau),Q(\omega,\tau))+2\int\int\left(\sqrt{P(\tau)}-\sqrt{Q(\tau)}\right)^2 Q(\omega|\tau)d\omega d\tau\\
&\le 4d_{\mathrm{H}}^2(P(\omega,\tau),Q(\omega,\tau)).
\end{aligned}$$

□

Lemma A.5. *Let $x\in\mathbb{R}^d$ be a random vector from a series of distributions $\{\mathcal{D}^k\}_k$ and let $U_k = U_1+\sum_{j<k}\mathbb{E}_{x\sim\mathcal{D}^j}[xx^\top]$ with $U_1\succeq\lambda I$ for some positive constant $\lambda>0$. Assume that $\|x\|_2\le 1$ almost surely. Then,*

$$\sum_{k=1}^K \min\left(\mathbb{E}_{x\sim\mathcal{D}^k}\left[\|x\|_{U_k^{-1}}^2\right],R\right) \le (1+R)d\log(1+K/\lambda).$$

This is minor variation of the standard result from (Abbasi-Yadkori et al., 2011). Differently from their result, here, we need to establish the bound for the expected U_k. Hence their result is not directly applied here.

Proof. We follow the same technique of (Abbasi-Yadkori et al., 2011).

$$\begin{aligned}
\sum_{k=1}^K \min\left(\mathbb{E}_{x\sim\mathcal{D}^k}\left[\|x\|_{U_k^{-1}}^2\right],R\right) &\le (1+R)\sum_{k=1}^K\log\left(1+\mathbb{E}_{x\sim\mathcal{D}^k}\left[\|x\|_{U_k^{-1}}^2\right]\right)\\
&\overset{(a)}{=}(1+R)\sum_{k=1}^K\log\left(1+\mathtt{Tr}(\mathbb{E}_{x\sim\mathcal{D}^k}\left[xx^\top\right]U_k^{-1})\right)
\end{aligned}$$

$$
\begin{aligned}
&= (1+R)\sum_{k=1}^{K}\log\left(1+\texttt{Tr}((U_{k+1}-U_k)U_k^{-1})\right)\\
&\le (1+R)\sum_{k=1}^{K}\log\texttt{det}\left(I_d+(U_k^{-1/2}(U_{k+1}-U_k)U_k^{-1/2})\right)\\
&= (1+R)\sum_{k=1}^{K}\log\frac{\texttt{det}U_{k+1}}{\texttt{det}(U_k)} = (1+R)\log\frac{\texttt{det}(U_{K+1})}{\texttt{det}(U_1)}\\
&\le (1+R)d\log(1+K/\lambda),
\end{aligned}
$$

where (a) is due to the linearity of trace operators. □

Lemma A.6. *Let x_k be any sequence of vectors in $\mathbb{R}^d$ where **rank**$(\{x_k\}_k) = r < d$, and let $U_k = \lambda I + \sum_{j<k} x_j x_j^\top$. Then,*

$$\sum_{j<k}\|x_k\|^2_{U_k^{-1}} \le r.$$

Proof. Again, we can express $a^\top A^{-1}a = \texttt{Tr}(aa^\top A^{-1})$, and thus

$$
\begin{aligned}
\sum_{j<k}\texttt{Tr}(x_j x_j^\top U_k^{-1}) &= \texttt{Tr}\left((\textstyle\sum_{j<k} x_j x_j^\top)U_k^{-1}\right)\\
&= \texttt{Tr}\left(I-\left(I+\lambda^{-1}\textstyle\sum_{j<k}x_jx_j^\top\right)^{-1}\right)\le r,
\end{aligned}
$$

where the inequality holds since the matrix inside Tr is at most rank r with eigenvalues less than or equal to one. □

Lemma A.7. *For any vectors a, b and positive definite matrices A, B such that $A, B \succeq \lambda_0 I$, we have*

$$\|a\|_{A^{-1}} - \|b\|_{B^{-1}} \le \frac{1}{\sqrt{\lambda_0}}\|a-b\|_2 + \|b\|_{B^{-1}}\|A^{-1/2}(B-A)B^{-1/2}\|_2.$$

Proof. The proof follows by algebraic manipulations:

$$
\begin{aligned}
\|a\|_{A^{-1}} - \|b\|_{B^{-1}} &= \frac{\|a\|^2_{A^{-1}} - \|b\|^2_{B^{-1}}}{\|a\|_{A^{-1}} + \|b\|_{B^{-1}}}\\
&= \frac{a^\top A^{-1}(a-b) + (a-b)^\top B^{-1}b + a^\top A^{-1}(B-A)B^{-1}b}{\|a\|_{A^{-1}} + \|b\|_{B^{-1}}}\\
&\le \frac{\|a\|_{A^{-1}}\|a-b\|_{A^{-1}} + \|a-b\|_{B^{-1}}\|b\|_{B^{-1}} + a^\top A^{-1}(B-A)B^{-1}b}{\|a\|_{A^{-1}} + \|b\|_{B^{-1}}}\\
&\le \frac{1}{\sqrt{\lambda_0}}\|a-b\|_2 + \|b\|_{B^{-1}}\|A^{-1/2}(B-A)B^{-1/2}\|_2.
\end{aligned}
$$

□

B. Proof of Upper Bounds

We remind the reader some notations we frequently use in the appendix.

$$
\begin{aligned}
&B(o, s_{+1}|s,a) = \mathbb{I}\cdot\textbf{diag}([\mathbb{P}(o,s_{+1}|m,s,a)]_{m=1}^M)\cdot\mathbb{I}^\dagger,\\
&b_0 = \mathbb{I}w,\\
&\tau_t = (s_1,a_1,o_1,...,s_t,a_t),\\
&\omega_t = (o_t,s_{t+1},a_{t+1},...,o_H),\\
&\psi(\omega_t,\iota|s_t)^\top = \boldsymbol{e}_\iota^\top\cdot\Pi_{h=t}^H B(o_h,s_{h+1}|s_h,a_h),
\end{aligned}
$$

$$b(\tau_t) = \Pi_{h=1}^{t-1} B(o_h, s_{h+1}|s_h, a_h) b_0,$$
$$\pi(\tau_t) = \Pi_{h=1}^{t} \pi(a_h|s_1, ..., s_h),$$
$$\pi(\omega_t|\tau_t) = \Pi_{h=t+1}^{H} \pi(a_h|s_1, ..., s_h).$$

We frequently use a shorthand for a pair of observations, $x_t := (s_t, a_t)$ and $y_t := (o_t, s_{t+1})$.

B.1. Proof of Theorem 4.1

There are several analysis techniques available in previous work (*e.g.,* Liu et al. (2022); Uehara et al. (2022); Liu et al. (2023); Chen et al. (2022); Huang et al. (2023)). Among all the above great works, we find the recent analysis of Huang et al. (2023) as particularly well-suited for our setting, and thus we adopt their proof ideas.

By the choice of model selection in the confidence set, it is sufficient to bound the sum TV-distances since

$$\sum_{k=1}^{K} V_{\theta^*}^{\pi^*_{\texttt{blind}}} - V_{\theta^*}^{\pi^k} \leq \sum_{k=1}^{K} V_{\theta^k}^{\pi_k} - V_{\theta^*}^{\pi^k} \leq H \cdot \sum_{k=1}^{K} d_{\texttt{TV}}(\mathbb{P}_{\theta^k}^{\pi^k}, \mathbb{P}_{\theta^*}^{\pi^k}).$$

At each episode $k \in [K]$, we start by unfolding the upper bound of the total-variation distance:

$$\begin{aligned} d_{\texttt{TV}}(\mathbb{P}_{\theta^*}^{\pi^k}(\tau, \iota), \mathbb{P}_{\theta}^{\pi^k}(\tau, \iota)) &\leq \sum_{\tau,\iota} \sum_{t=1}^{H} \pi(\tau) \cdot \left| \psi_{\theta^k}(\omega_{t+1}, \iota|x_{t+1})^\top b_{\theta^*}(\tau_{t+1}) - \psi_{\theta^k}(\omega_t, \iota|x_t)^\top b_{\theta^*}(\tau_t) \right| \\ &= \sum_{t=1}^{H} \sum_{\tau,\iota} \pi(\tau) \cdot \left| \psi_{\theta^k}(\omega_{t+1}, \iota|x_{t+1})^\top \left(B_{\theta^k}(y_t|x_t) - B_{\theta^*}(y_t|x_t) \right) b_{\theta^*}(\tau_t) \right|. \end{aligned}$$

We focus on bounding the inner summation fixing t. Every trajectory τ can be decomposed into τ_t and ω_t, and thus

$$\begin{aligned} &d_{\texttt{TV}}(\mathbb{P}_{\theta^*}^{\pi^k}(\tau, \iota), \mathbb{P}_{\theta}^{\pi^k}(\tau, \iota)) \\ &\leq \sum_{t} \sum_{\tau_t} \pi^k(\tau_t) \sum_{\omega_t,\iota} \pi^k(\omega_t|\tau_t) \cdot \left| \psi_{\theta^k}(\omega_{t+1}, \iota|x_{t+1})^\top \left(B_{\theta^k}(y_t|x_t) - B_{\theta^*}(y_t|x_t) \right) \mathbb{I}_{\theta^*} \mathbb{I}_{\theta^*}^{\dagger} b_{\theta^*}(\tau_t) \right|, \end{aligned} \tag{11}$$

where we used $\mathbb{I}_{\theta^*} \mathbb{I}_{\theta^*}^{\dagger} b_{\theta^*}(\cdot) = b_{\theta^*}(\cdot)$ since $b_{\theta^*}(\cdot)$ is in the column span of $\mathbb{I}_{\theta^*}$. Define

$$v_{\theta^*}(\tau_t) = \mathbb{I}_{\theta^*}^{\dagger} b_{\theta^*}(\tau_t), \text{ and }, \bar{v}_{\theta^*}(\tau_t) = \frac{v_{\theta^*}(\tau_t)}{\|v_{\theta^*}(\tau_t)\|_1},$$

which are the internal unnormalized and normalized latent belief states, respectively. Then the RHS in equation (11) can be expressed as

$$\sum_{\tau_t} \pi^k(\tau_t) \|\bar{v}_{\theta^*}(\tau_t)\|_1 \sum_{\omega_t,\iota} \pi^k(\omega_t|\tau_t) \cdot \left| \psi_{\theta^k}(\omega_{t+1}, \iota|x_{t+1})^\top \left(B_{\theta^k}(y_t|x_t) - B_{\theta^*}(y_t|x_t) \right) \mathbb{I}_{\theta^*} \bar{v}_{\theta^*}(\tau_t) \right|.$$

Define an elliptical potential matrix $\Lambda_*^k(s, a)$ as

$$\Lambda_*^k(s, a) = \lambda^* I + \sum_{j<k} \mathbb{E}_{\theta^*}^{\pi^j} \left[\mathbb{1}\left\{ (s_t, a_t) = (s, a) \right\} \bar{v}_{\theta^*}(\tau_t) \bar{v}_{\theta^*}(\tau_t)^\top \right],$$

where we define λ^* later (here, the choice of λ^* does not matter much). Using Cauchy-Schwartz inequality, we can separate the concentration argument and the pigeon-hole (a.k.a. elliptical potential lemma) argument. For simplicity, let $f(\omega_t, \iota) := \psi_{\theta^k}(\omega_{t+1}, \iota|x_{t+1})^\top \left(B_{\theta^k}(y_t|x_t) - B_{\theta^*}(y_t|x_t) \right) \mathbb{I}_{\theta^*}$. Then

$$\begin{aligned} &\sum_{\omega_t,\iota} \pi^k(\omega_t|\tau_t) \cdot \left| \psi_{\theta^k}(\omega_{t+1}, \iota|x_{t+1})^\top \left(B_{\theta^k}(y_t|x_t) - B_{\theta^*}(y_t|x_t) \right) \mathbb{I}_{\theta^*} \bar{v}_{\theta^*}(\tau_t) \right| \\ &= \sum_{\omega_t,\iota} \pi^k(\omega_t|\tau_t) \cdot |f(\omega_t, \iota) \bar{v}_{\theta^*}(\tau_t)| = \sum_{\omega_t,\iota} \pi^k(\omega_t|\tau_t) \cdot f(\omega_t, \iota) \texttt{sgn}(f(\omega_t, \iota) \bar{v}_{\theta^*}(\tau_t)) \cdot \bar{v}_{\theta^*}(\tau_t) \end{aligned}$$

$$\leq \left\| \sum_{\omega_t,\iota} \pi^k(\omega_t|\tau_t) \cdot f(\omega_t,\iota) \mathtt{sgn}(f(\omega_t,\iota)\bar{v}_{\theta^*}(\tau_t)) \right\|_{\Lambda_*^k(x_t)} \|\bar{v}_{\theta^*}(\tau_t)\|_{\Lambda_*^k(x_t)^{-1}} .$$

Checking the squared norm of the first part, we observe that

$$\begin{aligned}
&\left\| \sum_{\omega_t,\iota} \pi^k(\omega_t|\tau_t) \cdot f(\omega_t,\iota) \mathtt{sgn}(f(\omega_t,\iota)\bar{v}_{\theta^*}(\tau_t)) \right\|^2_{\Lambda_*^k(x_t)} \\
&= \underbrace{\lambda^* \left\| \sum_{\omega_t,\iota} \pi(\omega_t|\tau_t) f(\omega_t,\iota) \cdot \mathtt{sgn}(f(\omega_t,\iota)\bar{v}_{\theta^*}(\tau_t)) \right\|_2^2}_{(i)} \\
&\quad + \underbrace{\sum_{j<k} \mathbb{E}_{\theta^*}^{\pi^j} \left[\mathbb{1}\left\{ x_t^j = x_t \right\} \left(\sum_{\omega_t,\iota} \pi(\omega_t|\tau_t)(f(\omega_t,\iota)\bar{v}_{\theta^*}(\tau_t^j)) \cdot \mathtt{sgn}(f(\omega_t,\iota)\bar{v}_{\theta^*}(\tau_t)) \right)^2 \right]}_{(ii)} .
\end{aligned}$$

Bounding (i)**.** For any $m \in [M]$ we observe that

$$\begin{aligned}
&\left| \sum_{\omega_t,\iota} \pi(\omega_t|\tau_t) f(\omega_t,\iota) \boldsymbol{e}_m \cdot \mathtt{sgn}(f(\omega_t,\iota)\bar{v}_{\theta^*}(\tau_t)) \right| \leq \sum_{\omega_t,\iota} |\pi(\omega_t|\tau_t) f(\omega_t,\iota) \boldsymbol{e}_m| \\
&\leq \sum_{\omega_t,\iota} \pi(\omega_t|\tau_t) \left| \psi_{\theta^k}(\omega_{t+1},\iota|x_{t+1})^\top \left(B_{\theta^k}(y_t|x_t) - B_{\theta^*}(y_t|x_t) \right) \mathbb{I}_{\theta^*} \boldsymbol{e}_m \right| \\
&\leq \sum_{\omega_t,\iota} \pi(\omega_t|\tau_t) \left| \psi_{\theta^k}(\omega_t,\iota|x_t)^\top \mathbb{I}_{\theta^*} \boldsymbol{e}_m - \psi_{\theta^k}(\omega_{t+1},\iota|x_{t+1})^\top \mathbb{I}_{\theta^*} \boldsymbol{e}_m \cdot \mathbb{P}_{\theta^*}(y_t|m,x_t) \right| \\
&\leq \frac{2M}{\alpha} \|\mathbb{I}_\theta \boldsymbol{e}_m\|_1 = \frac{2M}{\alpha}.
\end{aligned}$$

Therefore, $(i) \leq \lambda^* M (2M/\alpha)^2 = 4M^3\lambda^*/\alpha^2$.

Bounding (ii)**.** Observe that

$$\begin{aligned}
&\sum_{\omega_t,\iota} \pi(\omega_t|\tau_t)(f(\omega_t,\iota)\bar{v}_{\theta^*}(\tau_t^j)) \cdot \mathtt{sgn}(f(\omega_t,\iota)\bar{v}_{\theta^*}(\tau_t)) \\
&\leq \sum_{\omega_t,\iota} \pi(\omega_t|\tau_t) \left| \psi_{\theta^k}(\omega_{t+1},\iota|x_{t+1})^\top \left(B_{\theta^k}(y_t|x_t) - B_{\theta^*}(y_t|x_t) \right) \mathbb{I}_{\theta^*} \bar{v}_{\theta^*}(\tau_t^j) \right| \\
&\leq \sum_{\omega_t,\iota} \pi(\omega_t|\tau_t) \left| \psi_{\theta^k}(\omega_{t+1},\iota|x_{t+1})^\top \left(B_{\theta^k}(y_t|x_t)\bar{b}_{\theta^k}(\tau_t^j) - B_{\theta^*}(y_t|x_t)\bar{b}_{\theta^*}(\tau_t^j) \right) \right| \\
&\quad + \sum_{\omega_t,\iota} \pi(\omega_t|\tau_t) \left| \psi_{\theta^k}(\omega_t,\iota|x_t)^\top \left(\bar{b}_{\theta^k}(\tau_t^j) - \bar{b}_{\theta^*}(\tau_t^j) \right) \right| \\
&\leq \frac{M}{\alpha} \left(\|\bar{b}_{\theta^k}(\tau_t^j) - \bar{b}_{\theta^*}(\tau_t^j)\|_1 + \sum_{y_t} \|B_{\theta^k}(y_t|x_t)\bar{b}_{\theta^k}(\tau_t^j) - B_{\theta^*}(y_t|x_t)\bar{b}_{\theta^*}(\tau_t^j)\|_1 \right),
\end{aligned}$$

where we denoted $\bar{b}_\theta = \mathbb{I}_\theta \bar{v}_\theta$ for any θ. The last inequality follows from the well-conditionedness of the system following equation (3). Then the statistical meaning of each term is given by

$$\begin{aligned}
&\boldsymbol{e}_\iota^\top \bar{b}_\theta(\tau_t^j) = \mathbb{P}_\theta^{\pi^j}(\iota|\tau_t^j), \\
&\mathbb{1}\left\{ x_t^j = x_t \right\} \boldsymbol{e}_\iota^\top B_\theta(y_t|x_t)\bar{b}_\theta(\tau_t^j) = \mathbb{P}_\theta^{\pi_j}(\iota, y_t|\tau_t^j).
\end{aligned}$$

The second equality can be verified by the following steps:

$$\begin{aligned}
&\mathbb{1}\left\{x_t^j = x_t\right\} \cdot \boldsymbol{e}_\iota^\top B_\theta(y_t|x_t)\bar{b}_\theta(\tau_t^j) \\
&= \frac{\mathbf{1}^\top \mathbf{diag}(\mathbb{P}_\theta(\iota|m))\Pi_{h=1}^t \mathbf{diag}(\mathbb{P}_\theta(y_h|m,x_h^j))w}{\|\bar{b}_\theta(\tau_t^j)\|_1} \\
&= \frac{\mathbf{1}^\top \mathbf{diag}(\mathbb{P}_\theta(\iota|m))\Pi_{h=1}^t \mathbf{diag}(\mathbb{P}_\theta(y_h|m,x_h^j))w}{\sum_{\iota'} \mathbf{1}^\top \mathbf{diag}(\mathbb{P}_\theta(\iota'|m))\Pi_{h=1}^{t-1} \mathbf{diag}_\theta(\mathbb{P}(y_h|m,x_h^j))w} \\
&= \frac{\pi^j(\tau_t^j)\mathbf{1}^\top \mathbf{diag}(\mathbb{P}_\theta(\iota|m))\Pi_{h=1}^t \mathbf{diag}(\mathbb{P}_\theta(y_h|m,x_h^j))w}{\sum_{\iota'} \pi^j(\tau_t^j)\mathbf{1}^\top \mathbf{diag}(\mathbb{P}_\theta(\iota'|m))\Pi_{h=1}^{t-1} \mathbf{diag}_\theta(\mathbb{P}(y_h|m,x_h^j))w} \\
&= \frac{\mathbb{P}_\theta^{\pi^j}(\iota, y_t, \tau_t^j)}{\mathbb{P}_\theta^{\pi^j}(\tau_t^j)} = \mathbb{P}_\theta^{\pi^j}(i, y_t|\tau_t^j).
\end{aligned}$$

In summary, we have $(ii) \le \frac{4M^2}{\alpha^2}\left(d_{\mathrm{TV}}^2(\mathbb{P}_{\theta^k}^{\pi^j}(\iota, y_t|\tau_t^j), \mathbb{P}_{\theta^*}^{\pi^j}(\iota, y_t|\tau_t^j))\right)$.

Combining bounds for (i) and (ii). Therefore, we can conclude that

$$\begin{aligned}
&\left\|\sum_{\omega_t,\iota} \pi^k(\omega_t|\tau_t) \cdot f(\omega_t,\iota)\mathtt{sgn}(f(\omega_t,\iota)\bar{v}_{\theta^*}(\tau_t))\right\|_{\Lambda_*^k(x_t)}^2 \\
&\le \frac{4M^3\lambda^*}{\alpha^2} + \frac{4M^2}{\alpha^2}\sum_{j<k} \mathbb{E}_{\theta^*}^{\pi^j}\left[d_{\mathrm{TV}}^2(\mathbb{P}_{\theta^k}^{\pi^j}(\iota, y_t|\tau_t^j), \mathbb{P}_{\theta^*}^{\pi^j}(\iota, y_t|\tau_t^j))\right] \\
&\le \frac{4M^3\lambda^*}{\alpha^2} + \frac{8M^2}{\alpha^2}\sum_{j<k} \mathbb{E}_{\theta^*}^{\pi^j}\left[d_{\mathrm{H}}^2(\mathbb{P}_{\theta^k}^{\pi^j}(\iota, y_t|\tau_t^j), \mathbb{P}_{\theta^*}^{\pi^j}(\iota, y_t|\tau_t^j))\right] \\
&\le \frac{4M^3\lambda^*}{\alpha^2} + \frac{32M^2}{\alpha^2}\sum_{j<k} d_{\mathrm{H}}^2(\mathbb{P}_{\theta^k}^{\pi^j}(\iota, y_t, \tau_t^j), \mathbb{P}_{\theta^*}^{\pi^j}(\iota, y_t, \tau_t^j)),
\end{aligned}$$

where we used Lemma A.4. Finally, due to the concentration of the square sum of Helligner distances (Lemma A.2), we can conclude that

$$\left\|\sum_{\omega_t,\iota} \pi^k(\omega_t|\tau_t) \cdot f(\omega_t,\iota)\mathtt{sgn}(f(\omega_t,\iota)\bar{v}_{\theta^*}(\tau_t))\right\|_{\Lambda_*^k(x_t)}^2 \lesssim \frac{M^2}{\alpha^2}(\lambda^* M + \beta).$$

Plugging this bound back to equation (11), we have

$$\begin{aligned}
d_{\mathrm{TV}}(\mathbb{P}_{\theta^*}^{\pi^k}(\tau,\iota), \mathbb{P}_{\theta^k}^{\pi^k}(\tau,\iota)) &\lesssim \frac{M}{\alpha}\sqrt{(\lambda^* M + \beta)} \cdot \sum_t \sum_{\tau_t} \pi^k(\tau_t)\|\bar{v}_{\theta^*}(\tau_t)\|_1 \|\bar{v}_{\theta^*}(\tau_t)\|_{\Lambda_*^k(x_t)^{-1}} \\
&= \frac{M}{\alpha}\sqrt{(\lambda^* M + \beta)} \cdot \sum_t \mathbb{E}_{\theta^*}^{\pi^k}\left[\|\bar{v}_{\theta^*}(\tau_t)\|_{\Lambda_*^k(x_t)^{-1}}\right].
\end{aligned}$$

Finally, summing up over all episodes, we have

$$\begin{aligned}
\sum_{k=1}^K d_{\mathrm{TV}}(\mathbb{P}_{\theta^*}^{\pi^k}(\tau,\iota), \mathbb{P}_{\theta^k}^{\pi^k}(\tau,\iota)) &\lesssim \frac{M}{\alpha}\sqrt{(\lambda^* M + \beta)} \cdot \sum_{t=1}^H \sum_{k=1}^K \mathbb{E}_{\theta^*}^{\pi^k}\left[\|\bar{v}_{\theta^*}(\tau_t)\|_{\Lambda_*^k(x_t)^{-1}}\right] \\
&\le \frac{M}{\alpha}\sqrt{(\lambda^* M + \beta)K} \cdot \sum_{t=1}^H \sqrt{\sum_{k=1}^K \mathbb{E}_{\theta^*}^{\pi^k}\left[\|\bar{v}_{\theta^*}(\tau_t)\|_{\Lambda_*^k(x_t)^{-1}}^2\right]}.
\end{aligned}$$

Applying the expectation version of the elliptical potential lemma (see Lemma A.5), by considering $\bar{v}_{\theta^*}(\tau_t)$ in the space of $\mathbb{R}^{MSA}$, and setting $\lambda^* = O(1)$, $\beta = \log(K|\Theta|/\delta) > M$, we have

$$\sum_{k=1}^{K} d_{\text{TV}}(\mathbb{P}_{\theta^*}^{\pi^k}(\tau,\iota), \mathbb{P}_{\theta^k}^{\pi^k}(\tau,\iota)) \lesssim \frac{MH}{\alpha}\sqrt{MSAK\beta\log(K)}, \tag{12}$$

with probability at least $1-\delta$. Consequently, the regret bound is given by

$$\sum_{k=1}^{K} V_{\theta^*}^{\pi^*_{\texttt{blind}}} - V_{\theta^*}^{\pi^k} \lesssim \frac{M^{3/2}H^2}{\alpha}\sqrt{SAK\log(K|\Theta|/\delta)\log(K)},$$

completing the proof.

B.2. Proof of Lemma 5.1

Proof. Recall that

$$\begin{aligned}\pi(\omega_t)\psi(\omega_t,\iota|x_t)^\top &= \pi(\omega_t)\cdot \boldsymbol{e}_\iota^\top B(y_H|x_H)...B(y_t|x_t)\\ &= \mathbb{I}(\iota)^\top \mathbf{diag}(\mathbb{P}^\pi(\omega_t|m,x_t))\mathbb{I}^\dagger.\end{aligned}$$

Thus,

$$\begin{aligned}\sum_{\omega_t}\pi(\omega_t)|\psi(\omega_t,\iota|x_t)^\top b| &= \sum_{\omega_t}|\mathbb{I}(\iota)^\top \mathbf{diag}(\mathbb{P}^\pi(\omega_t|m,x_t))\mathbb{I}^\dagger b|\\ &\le \sum_{\omega_t}\sum_m |\mathbb{P}(\iota|m)\mathbb{P}^\pi(\omega_t|m,x_t)|\cdot|\boldsymbol{e}_m^\top\mathbb{I}^\dagger b|\\ &\le \sum_m \mathbb{P}(\iota|m)|\boldsymbol{e}_m^\top\mathbb{I}^\dagger b| \le \|\mathbb{I}(\iota)\|_\infty\|\mathbb{I}^\dagger b\|_1.\end{aligned}$$

Now applying Lemma G.4 in (Liu et al., 2023), there exists a left-inverse of $\mathbb{I}$ such that $\|\mathbb{I}^\dagger b\|_1 \le M\|b\|_1/\alpha$, and we have the result. □

B.3. Proof of Theorem 4.4

We divide the proof of this theorem into two parts. In the first part, we prove the required number of episodes until Algorithm 2 terminates. In the second part, we show the optimality of the returned model in a larger class of prospective side information exploiting policies Π.

B.3.1. Proof Part I

The first part largely follows the proofs in Huang et al. (2023) for the reward-free exploration until the sum of trajectory bonuses becomes small. The key step is connecting the trajectory bonuses between two different models in the confidence set. Define the bonus counterpart in the true environment:

$$\begin{aligned}\Lambda_t^k(x) &= \lambda_0 I + \sum_{j<k}\mathbb{1}\left\{x_t^j = x\right\}\bar{b}_{\theta^*}(\tau_t^j)\bar{b}_{\theta^*}(\tau_t^j)^\top,\\ \tilde{r}_*^k(\tau_t) &= \|\bar{b}_{\theta^*}(\tau_t)\|_{\Lambda_t^k(x_t)^{-1}}.\end{aligned}$$

Then we compare that

$$\|\bar{b}_{\theta^k}(\tau_t)\|_{\hat{\Lambda}_t^k(x_t)^{-1}} - \|\bar{b}_{\theta^*}(\tau_t)\|_{\Lambda_t^k(x_t)^{-1}}.$$

Using Lemma A.7, we can show that

$$\|\bar{b}_{\theta^k}(\tau_t)\|_{\hat{\Lambda}_t^k(x_t)^{-1}} - \|\bar{b}_{\theta^*}(\tau_t)\|_{\Lambda_t^k(x_t)^{-1}}$$

$$\leq \frac{1}{\sqrt{\lambda_0}}\|\bar{b}_{\theta^k}(\tau_t) - \bar{b}_{\theta^*}(\tau_t)\|_2 + \|\bar{b}_{\theta^*}(\tau_t)\|_{\Lambda_t^k(x_t)^{-1}} \underbrace{\left\|\sum_{j<k} \mathbb{1}\left\{x_t^j = x_t\right\} \hat{\Lambda}_t^k(x_t)^{-1/2}(\bar{b}_{\theta^*}(\tau_t^j)\bar{b}_{\theta^*}(\tau_t^j)^\top - \bar{b}_{\theta^k}(\tau_t^j)\bar{b}_{\theta^k}(\tau_t^j)^\top)\Lambda_t^k(x_t)^{-1/2}\right\|_2}_{(a)}.$$

(a) can be further bounded by

$$\begin{aligned}
(a) &\leq \max_{u,v:\|u\|_2=1,\|v\|_2=1} \sum_{j<k} \mathbb{1}\left\{x_t^j = x_t\right\} u\hat{\Lambda}_t^k(x_t)^{-1/2}(\bar{b}_{\theta^*}(\tau_t^j)\bar{b}_{\theta^*}(\tau_t^j)^\top - \bar{b}_{\theta^k}(\tau_t^j)\bar{b}_{\theta^k}(\tau_t^j)^\top)\Lambda_t^k(x_t)^{-1/2}v \\
&\leq \max_{u,v:\|u\|_2=1,\|v\|_2=1} \sum_{j<k} \mathbb{1}\left\{x_t^j = x_t\right\} \left|u\hat{\Lambda}_t^k(x_t)^{-1/2}\bar{b}_{\theta^k}(\tau_t^j)\right| \left|(\bar{b}_{\theta^*}(\tau_t^j)^\top - \bar{b}_{\theta^k}(\tau_t^j))^\top\Lambda_t^k(x_t)^{-1/2}v\right| \\
&+ \max_{u,v:\|u\|_2=1,\|v\|_2=1} \sum_{j<k} \mathbb{1}\left\{x_t^j = x_t\right\} \left|u\hat{\Lambda}_t^k(x_t)^{-1/2}(\bar{b}_{\theta^*}(\tau_t^j) - \bar{b}_{\theta^k}(\tau_t^j))\right| \left|\bar{b}_{\theta^*}(\tau_t^j)^\top\Lambda_t^k(x_t)^{-1/2}v\right| \\
&\leq \sqrt{\sum_{j<k} \mathbb{1}\left\{x_t^j = x_t\right\} \|\bar{b}_{\theta^k}(\tau_t^j)\|^2_{\hat{\Lambda}_t^k(x_t)^{-1}}} \sqrt{\sum_{j<k} \mathbb{1}\left\{x_t^j = x_t\right\} \|\bar{b}_{\theta^*}(\tau_t^j) - \bar{b}_{\theta^k}(\tau_t^j)\|^2_{\Lambda_t^k(x_t)^{-1}}} \\
&+ \sqrt{\sum_{j<k} \mathbb{1}\left\{x_t^j = x_t\right\} \|\bar{b}_{\theta^*}(\tau_t^j)\|^2_{\Lambda_t^k(x_t)^{-1}}} \sqrt{\sum_{j<k} \mathbb{1}\left\{x_t^j = x_t\right\} \|\bar{b}_{\theta^*}(\tau_t^j) - \bar{b}_{\theta^k}(\tau_t^j)\|^2_{\hat{\Lambda}_t^k(x_t)^{-1}}} \\
&\overset{(b)}{\leq} \sqrt{\frac{M}{\lambda_0}} \sqrt{\sum_{j<k} \|\bar{b}_{\theta^*}(\tau_t^j) - \bar{b}_{\theta^k}(\tau_t^j)\|_2^2} \leq \sqrt{\frac{M}{\lambda_0}} \sqrt{\sum_{j<k} d_{\mathrm{TV}}^2(\mathbb{P}_{\theta^*}^{\pi_j}(\iota|\tau_t^j), \mathbb{P}_{\theta^k}^{\pi_j}(\iota|\tau_t^j))} \lesssim \sqrt{\frac{M\beta}{\lambda_0}},
\end{aligned}$$

where for (b), we used Lemma A.6.

Now taking expectation on both sides, we have

$$\mathbb{E}_{\theta^*}^{\pi^k}\left[\|\bar{b}_{\theta^k}(\tau_t)\|_{\hat{\Lambda}_t^k(x_t)^{-1}}\right] \leq \left(1 + O(1)\cdot\sqrt{M\beta/\lambda_0}\right)\mathbb{E}_{\theta^*}^{\pi^k}\left[\|\bar{b}_{\theta^*}(\tau_t)\|_{\Lambda_t^k(x_t)^{-1}}\right] + \frac{O(1)}{\sqrt{\lambda_0}} d_{\mathrm{TV}}(\mathbb{P}_{\theta^*}^{\pi_k}, \mathbb{P}_{\theta^k}^{\pi_k}),$$

where we used

$$\begin{aligned}
\mathbb{E}_{\theta^*}^{\pi^k}\left[\|\bar{b}_{\theta^k}(\tau_t) - \bar{b}_{\theta^*}(\tau_t)\|_2\right] &\leq \mathbb{E}_{\theta^*}^{\pi^k}\left[\|\bar{b}_{\theta^k}(\tau_t) - \bar{b}_{\theta^*}(\tau_t)\|_1\right] \\
&\leq \mathbb{E}_{\theta^*}^{\pi^k}\left[d_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\pi^k}(\iota|\tau_t), \mathbb{P}_{\theta^*}^{\pi^k}(\iota|\tau_t)\right)\right] \\
&\leq 2 d_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\pi^k}(\iota, \tau_t), \mathbb{P}_{\theta^*}^{\pi^k}(\iota, \tau_t)\right).
\end{aligned}$$

To proceed, note that $\|\bar{b}_{\theta^k}(\tau_t)\|_{\hat{\Lambda}_t^k(x_t)^{-1}} \leq \frac{1}{\sqrt{\lambda_0}}$ almost surely, and thus,

$$\mathbb{E}_{\theta^k}^{\pi^k}\left[\|\bar{b}_{\theta^k}(\tau_t)\|_{\hat{\Lambda}_t^k(x_t)^{-1}}\right] \leq \mathbb{E}_{\theta^*}^{\pi^k}\left[\|\bar{b}_{\theta^k}(\tau_t)\|_{\hat{\Lambda}_t^k(x_t)^{-1}}\right] + \frac{1}{\sqrt{\lambda_0}} d_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\pi^k}, \mathbb{P}_{\theta^*}^{\pi^k}\right).$$

Therefore, summing over K episodes, we have

$$\begin{aligned}
\sum_{k=1}^K \mathbb{E}_{\theta^k}^{\pi^k}\left[\|\bar{b}_{\theta^k}(\tau_t)\|_{\hat{\Lambda}_t^k(x_t)^{-1}}\right] &\leq \left(1 + O(1)\cdot\sqrt{M\beta/\lambda_0}\right)\sum_{k=1}^K \mathbb{E}_{\theta^*}^{\pi^k}\left[\|\bar{b}_{\theta^*}(\tau_t)\|_{\Lambda_t^k(x_t)^{-1}}\right] \\
&+ \frac{O(1)}{\sqrt{\lambda_0}} \sum_{k=1}^K d_{\mathrm{TV}}\left(\mathbb{P}_{\theta^k}^{\pi^k}, \mathbb{P}_{\theta^*}^{\pi^k}\right).
\end{aligned}$$

For the second term, we can apply equation (12). For the first term, we can first apply Azuma-Hoeffding inequality on

$$\sum_{k=1}^K \left(\mathbb{E}_{\theta^*}^{\pi^k}\left[\|\bar{b}_{\theta^*}(\tau_t)\|_{\Lambda_t^k(x_t)^{-1}}\right] - \|\bar{b}_{\theta^*}(\tau_t^k)\|_{\Lambda_t^k(x_t)^{-1}}\right),$$

and apply the empirical version of elliptical potential lemma (Lemma A.5). This gives

$$\sum_{k=1}^{K} \mathbb{E}_{\theta^k}^{\pi^k}\left[\|\bar{b}_{\theta^k}(\tau_t)\|_{\hat{\Lambda}_t^k(x_t)^{-1}}\right] \lesssim \sqrt{MSAK\log(K)}\left(1+O(1)\cdot\sqrt{M\beta/\lambda_0}+(MH/\alpha)\cdot\sqrt{\beta/\lambda_0}\right).$$

With the choice of $\lambda_0 = \frac{\beta M^2 H^2}{\alpha^2}$, the Algorithm 2 must terminate after at most K episodes where

$$K = O\left(\frac{MSA\log(K)}{\epsilon_{\mathrm{pe}}^2}\right).$$

B.3.2. Proof Part II

Now suppose Algorithm 2 terminated with the model θ that has the desired property:

$$\max_{\pi\in\Pi_{\mathtt{blind}}} V_{\theta,\tilde{r}}^{\pi} := \mathbb{E}_{\theta}^{\pi}\left[\textstyle\sum_t \|\bar{b}_{\theta}(\tau_t)\|_{\hat{\Lambda}_t^k(x_t)^{-1}}\right] \le \epsilon_{\mathrm{pe}}.$$

Assuming this event holds true we continue the proof.

Proof. We can express the total-variation distance between θ^* and θ as

$$d_{\mathrm{TV}}(\mathbb{P}_{\theta^*}^{\pi}(\tau,\iota),\mathbb{P}_{\theta}^{\pi}(\tau,\iota)) \le \sum_{t=1}^{H}\sum_{\tau}\pi(\tau)\cdot\left|\psi_{\theta^*}(\omega_{t+1},\iota|x_{t+1})^{\top}\left(B_{\theta^*}(y_t|x_t)-B_{\theta}(y_t|x_t)\right)b_{\theta}(\tau_t)\right|$$
$$\le \sum_{\iota}\sum_{t=1}^{H}\sum_{\tau_t}\pi(\tau_t|\iota)\sum_{\omega_t}\pi(\omega_t|\tau_t,\iota)\cdot\left|\psi_{\theta^*}(\omega_{t+1},\iota|x_{t+1})^{\top}\left(B_{\theta^*}(y_t|x_t)-B_{\theta}(y_t|x_t)\right)b_{\theta}(\tau_t)\right|.$$

Notice that this time, we use θ^* to express the future prediction, and θ to express the history part in the above equation. Now we fix ι, t and τ_t, and focus on bounding the inside summation. The first step is to normalize the belief state and rewrite the inner sum as:

$$\sum_{\tau_t}\pi(\tau_t|\iota)\sum_{\omega_t}\pi^k(\omega_t|\iota,\tau_t)\cdot\left|\psi_{\theta^*}(\omega_{t+1},\iota|x_{t+1})^{\top}\left(B_{\theta^*}(y_t|x_t)-B_{\theta}(y_t|x_t)\right)b_{\theta}(\tau_t)\right| \tag{13}$$
$$=\sum_{\tau_t}\pi(\tau_t|\iota)\|b_{\theta}(\tau_t)\|_1\sum_{\omega_t}\pi^k(\omega_t|\iota,\tau_t)\cdot\left|\psi_{\theta^*}(\omega_{t+1},\iota|x_{t+1})^{\top}\left(B_{\theta^*}(y_t|x_t)-B_{\theta}(y_t|x_t)\right)\bar{b}_{\theta}(\tau_t)\right|,$$

where $\bar{b}_{\theta}(\tau_t)=\frac{b_{\theta}(\tau_t)}{\|b_{\theta}(\tau_t)\|_1}$ are the normalized predictive representation of belief states. Then note that $\pi(\tau_t|\iota)\|b_{\theta}(\tau_t)\|_1 = \mathbb{P}_{\theta^*}^{\pi(\cdot|\iota)}(\tau_t)$, ***i.e.*, a marginalized probability of τ_t when running a prospective side information blind policy** $\pi(\cdot|\iota)$:

$$\mathbb{P}_{\theta^*}^{\pi(\cdot|\iota)}(\tau_t)=\sum_{\iota'}\mathbb{P}_{\theta^*}^{\pi(\cdot|\iota)}(\tau_t,\iota'),$$

as if we do not use the true prospective side information but instead use an arbitrary dummy variable ι to instantiate a blind policy. Thus, we can express (13) as

$$\mathbb{E}_{\tau_t\sim\mathbb{P}_{\theta}^{\pi(\cdot|\iota)}(\cdot)}\left[\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)\left|\psi_{\theta^*}(\omega_{t+1},\iota|x_{t+1})^{\top}\left(B_{\theta^*}(y_t|x_t)-B_{\theta}(y_t|x_t)\right)\bar{b}_{\theta}(\tau_t)\right|\right].$$

Recall the empirical pseudo-count matrix:

$$\hat{\Lambda}(s,a)=\lambda_0 I+\sum_{k\in[K]}\left[\mathbb{1}\left\{(s_t^k,a_t^k)=(s,a)\right\}\cdot\bar{b}_{\theta}(\tau_t^k)\bar{b}_{\theta}(\tau_t^k)^{\top}\right].$$

For simplicity, let $f(\omega_t):=\psi_{\theta^*}(\omega_{t+1},\iota|x_{t+1})^{\top}\left(B_{\theta^*}(y_t|x_t)-B_{\theta}(y_t|x_t)\right)$ (f is only a function of ω_t as other variables are fixed at this point). Using Cauchy-Schwartz inequality, we have

$$(13)\le\mathbb{E}_{\tau_t\sim\mathbb{P}_{\theta}^{\pi(\cdot|\iota)}}\left[\left\|\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)f(\omega_t)\cdot\mathtt{sgn}(f(\omega_t)^{\top}\bar{b}_{\theta}(\tau_t))\right\|_{\hat{\Lambda}(x_t)}\|\bar{b}_{\theta}(\tau_t)\|_{\hat{\Lambda}(x_t)^{-1}}\right].$$

To bound the concentration bound, we can check that

$$\begin{aligned}
&\left\|\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)f(\omega_t)\cdot \mathtt{sgn}(f(\omega_t)^\top \bar{b}_\theta(\tau_t))\right\|^2_{\hat{\Lambda}(x_t)}\\
&\le \lambda_0\left\|\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)f(\omega_t)\cdot \mathtt{sgn}(f(\omega_t)^\top \bar{b}_\theta(\tau_t))\right\|^2_2\\
&+\sum_{k\in[K]}\mathbb{1}\left\{x_t^k=x_t\right\}\left(\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)\left(f(\omega_t)^\top \bar{b}_\theta(\tau_t^k)\right)\cdot \mathtt{sgn}(f(\omega_t)^\top \bar{b}_\theta(\tau_t))\right)^2. \qquad (14)
\end{aligned}$$

For the term with λ_0, note that any vector v that lies on the orthogonal complement of the span of $\mathbb{I}_\theta$, $\mathbb{I}_\theta^\dagger v=0$. Consider a vector $v=\mathbb{I}_\theta u$ such that $\|\mathbb{I}_\theta u\|_2\le 1$. Note that to satisfy this condition, u cannot be too large: $\|u\|_1\le \max_{\|v\|_2=1}\|\mathbb{I}_\theta^\dagger v\|_1\le \max_{\|v\|_1=1}\|\mathbb{I}_\theta^\dagger v\|_1\le \frac{M}{\alpha}$. Thus,

$$\begin{aligned}
&\left|\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)f(w_t)v\right|\\
&\le\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)|\psi_{\theta^*}(\omega_{t+1},\iota|x_{t+1})^\top\left(B_{\theta^*}(y_t|x_t)-B_\theta(y_t|x_t)\right)\mathbb{I}_\theta u)|\\
&\le\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)|\psi_{\theta^*}(\omega_t,\iota|x_t)^\top\mathbb{I}_\theta u-\psi_{\theta^*}(\omega_{t+1},\iota|x_{t+1})^\top\mathbb{I}_\theta\mathbf{diag}(\mathbb{P}_\theta(y_t|m,x_t))u|\\
&\le\frac{2M}{\alpha}\|\mathbb{I}_{\theta^*}(\iota)\|_\infty\|u\|_1=\frac{2M^2}{\alpha^2}\|\mathbb{I}_{\theta^*}(\iota)\|_\infty,
\end{aligned}$$

where we applied Lemma A.1, and therefore

$$\left\|\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)f(w_t)\right\|_2^2\le\frac{4M^4}{\alpha^4}\|\mathbb{I}_{\theta^*}(\iota)\|_\infty^2.$$

To bound the second term in (14), first we observe the term inside the summation (over k) is only nonzero when $x_t^k=x_t$, i.e., $(s_t^k,a_t^k)=(s_t,a_t)$. We have that

$$\begin{aligned}
&\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)\left(f(\omega_t)^\top \bar{b}_\theta(\tau_t^k)\right)\cdot \mathtt{sgn}(f(\omega_t)^\top \bar{b}_\theta(\tau_t))\\
&\le\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)\left|f(\omega_t)^\top \bar{b}_\theta(\tau_t^k)\right|\\
&=\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)\left|\psi_{\theta^*}(\omega_{t+1},\iota|x_{t+1})^\top\left(B_{\theta^*}(y_t|x_t)-B_\theta(y_t|x_t)\right)\bar{b}_\theta(\tau_t^k)\right|\\
&\le\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)\left|\psi_{\theta^*}(\omega_{t+1},\iota|x_{t+1})^\top\left(B_{\theta^*}(y_t|x_t)\bar{b}_{\theta^*}(\tau_t^k)-B_\theta(y_t|x_t)\bar{b}_\theta(\tau_t^k)\right)\right|\\
&\quad+\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)\left|\psi_{\theta^*}(\omega_t,\iota|x_t)^\top\left(\bar{b}_{\theta^*}(\tau_t^k)-\bar{b}_\theta(\tau_t^k)\right)\right|\\
&\le\frac{M}{\alpha}\|\mathbb{I}_{\theta^*}(\iota)\|_\infty\cdot\left(\sum_{y_t}\|B_\theta(y_t|x_t)\bar{b}_\theta(\tau_t^k)-B_{\theta^*}(y_t|x_t)\bar{b}_{\theta^*}(\tau_t^k)\|_1+\|\bar{b}_\theta(\tau_t^k)-\bar{b}_{\theta^*}(\tau_t^k)\|_1\right),
\end{aligned}$$

where we denote $\bar{b}_\theta=\mathbb{I}_\theta\bar{b}_\theta$. We can check the meaning of each term: for any $\iota'\in\mathcal{I}$ and any *blind* policy $\pi\in\Pi_{\mathtt{blind}}$,

$$\begin{aligned}
&\mathbb{1}\left\{x_t^k=x_t\right\}\cdot\boldsymbol{e}_{\iota'}^\top B_\theta(y_t|x_t)\bar{b}_\theta(\tau_t^k)\\
&=\frac{\mathbb{1}^\top\mathbf{diag}(\mathbb{P}_\theta(\iota'|m))\Pi_{h=1}^t\mathbf{diag}(\mathbb{P}_\theta(y_h|m,x_h^k))w}{\|v_\theta(\tau_t^k)\|_1}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{\mathbb{1}^\top \mathbf{diag}(\mathbb{P}_\theta(\iota'|m))\Pi_{h=1}^{t}\mathbf{diag}(\mathbb{P}_\theta(y_h|m,x_h^k))w}{\sum_{\iota''}\mathbb{1}^\top \mathbf{diag}(\mathbb{P}_\theta(\iota''|m))\Pi_{h=1}^{t-1}\mathbf{diag}_\theta(\mathbb{P}(y_h|m,x_h^k))w} \\
&= \frac{\pi(\tau_t^k)\cdot\mathbb{1}^\top \mathbf{diag}(\mathbb{P}(\iota'|m))\Pi_{h=1}^{t}\mathbf{diag}(\mathbb{P}(y_h|m,x_h^k))w}{\sum_{\iota''}\pi(\tau_t^k)\cdot\mathbb{1}^\top \mathbf{diag}(\mathbb{P}(\iota''|m))\Pi_{h=1}^{t-1}\mathbf{diag}(\mathbb{P}(y_h|m,x_h^k))w} \\
&= \frac{\mathbb{P}_\theta^\pi(\iota',y_t,\tau_t^k)}{\mathbb{P}_\theta^\pi(\tau_t^k)} = \mathbb{P}_\theta^\pi(i',y_t|\tau_t^k).
\end{aligned}
$$

To proceed, let the prospective side information blind policy executed on the k^{th} episode be π^k. We have that

$$
\begin{aligned}
&\sum_{\omega_t}\mathbb{1}\left\{x_t^k=x_t\right\}\pi(\omega_t|\iota,\tau_t)\left(f(\omega_t)^\top\bar{b}_\theta(\tau_t^k)\right)\cdot\mathtt{sgn}(f(\omega_t)^\top\bar{b}_\theta(\tau_t)) \\
&\le \frac{2M}{\alpha}\|\mathbb{I}_{\theta^*}(\iota)\|_\infty\mathbb{1}\left\{x_t^k=x_t\right\}\cdot d_{\mathrm{TV}}\left(\mathbb{P}_{\theta^*}^{\pi^k}(\iota',y_t|\tau_t^k),\mathbb{P}_\theta^{\pi^k}(\iota',y_t|\tau_t^k)\right).
\end{aligned}
$$

Combining the result, we conclude that

$$
\begin{aligned}
(13) &\le \mathbb{E}_{\tau_t\sim\mathbb{P}_\theta^{\pi(\cdot|\iota)}}\left[\left\|\sum_{\omega_t}\pi(\omega_t|\iota,\tau_t)f(\omega_t)\cdot\mathtt{sgn}(f(\omega_t)^\top\bar{b}_\theta(\tau_t))\right\|_{\hat{\Lambda}(x_t)}\|\bar{b}_\theta(\tau_t)\|_{\hat{\Lambda}(x_t)^{-1}}\right] \\
&\le \mathbb{E}_{\tau_t\sim\mathbb{P}_\theta^{\pi(\cdot|\iota)}(\cdot)}\left[\frac{2M}{\alpha}\|\mathbb{I}_{\theta^*}(\iota)\|_\infty\cdot c(x_t)\|\bar{b}_\theta(\tau_t)\|_{\hat{\Lambda}(x_t)^{-1}}\right],
\end{aligned}
$$

where

$$
\begin{aligned}
c(x_t) &= \sqrt{\frac{M^2}{\alpha^2}\lambda_0+\sum_{k\in[K]}\left[\mathbb{1}\left\{x_t^k=x_t\right\}d_{\mathrm{TV}}^2\left(\mathbb{P}_{\theta^*}^{\pi^k}(\iota',y_t'|\tau_t^k),\mathbb{P}_\theta^{\pi^k}(\iota',y_t'|\tau_t^k)\right)\right]} \\
&\le \sqrt{\frac{M^2}{\alpha^2}\lambda_0+\sum_{k\in[K]}d_{\mathrm{TV}}^2\left(\mathbb{P}_{\theta^*}^{\pi^k}(\iota',y_t'|\tau_t^k),\mathbb{P}_\theta^{\pi^k}(\iota',y_t'|\tau_t^k)\right)} \\
&\lesssim \sqrt{\frac{M^2}{\alpha^2}\lambda_0+\beta} := c_{\max},
\end{aligned}
$$

where in the second inequality, we used Lemma A.3. Now proceeding,

$$
\begin{aligned}
&\sum_{\iota,\tau_H}|\mathbb{P}_{\theta^*}^\pi(\iota,\tau_H)-\mathbb{P}_\theta^\pi(\iota,\tau_H)| \\
&\le \frac{2M}{\alpha}c_{\max}\sum_\iota\|\mathbb{I}_{\theta^*}(\iota)\|_\infty\sum_t\mathbb{E}_{\tau_t\sim\mathbb{P}_\theta^{\pi(\cdot|\iota)}}\left[\|\bar{b}_\theta(\tau_t)\|_{\hat{\Lambda}(x_t)^{-1}}\right] \\
&= \frac{2M}{\alpha}c_{\max}\sum_\iota\|\mathbb{I}_{\theta^*}(\iota)\|_\infty\cdot\mathbb{E}_\theta^{\pi(\cdot|\iota)}\left[\sum_t\|\bar{b}_\theta(\tau_t)\|_{\hat{\Lambda}(x_t)^{-1}}\right] \\
&\le \frac{2M}{\alpha}c_{\max}\sum_\iota\|\mathbb{I}_{\theta^*}(\iota)\|_\infty\cdot\max_{\pi\in\Pi_{\mathrm{blind}}}\mathbb{E}_\theta^\pi\left[\sum_t\|\bar{b}_\theta(\tau_t)\|_{\hat{\Lambda}(x_t)^{-1}}\right] \\
&\overset{(a)}{\le} \frac{2M^2}{\alpha}\sqrt{\frac{M^2\lambda_0}{\alpha^2}+\beta}\cdot\epsilon_{\mathtt{pe}},
\end{aligned}
$$

where (a) comes from $\sum_\iota\|\mathbb{I}_{\theta^*}(\iota)\|_\infty\le\sum_m\sum_\iota\mathbb{I}_{\theta^*}(m,\iota)=M$. With the choice of $\lambda_0=\beta M^2H^2/\alpha^2$, by setting $\epsilon_{\mathtt{pe}}:=\frac{\alpha\epsilon}{10HM^2\sqrt{M^4H^2\beta/\alpha^4}}$ ensures that

$$
|V_\theta^\pi-V_{\theta^*}^\pi|\le Hd_{\mathrm{TV}}(\mathbb{P}_\theta^\pi,\mathbb{P}_{\theta^*}^\pi)\le\epsilon/2,
$$

for all $\pi\in\Pi$. Therefore, optimizing over θ gives ϵ-optimal policy for θ^*, completing the proof. □

C. Lower Bound Proofs

We first complete the construction of the hard instance family deferred from the main text. Recall that we defined:

1. Action space:
   - $\mathcal{A}_{\text{exploit}} = \{a^m_{\text{exploit}}\}$ for $m = M/2+1, ..., M$
   - $\mathcal{A}_{\text{explore}}$: contains the true exploration action (at the initial step) a^*_{explore} and dummy actions
   - $\mathcal{A}_{\text{control}}$: contains the optimal actions a^*_t for $t \in [d]$ and dummy actions
2. State space:
   - $s_{\text{init}}, s_{\text{ter}}$: initial and terminated state
   - $s^{\text{hard}}_{1:d}$: chained states in the hard-to-learn chain
   - $s^{\text{ref}}_{1:d}$: chained states in the reference chain
3. MDP groups in an LMDP:
   - $\mathcal{G}_{\text{learn}}$: a set of MDPs that needs to be acted optimally in the hard-to-learn chain
   - $\mathcal{G}_{\text{ref}}$: a set of MDPs that confuses the identity of hard-to-learn and reference chains
   - $\mathcal{G}_{\text{obs}}$: a set of MDPs whose identity is strongly correlated to the prospective side information

Initial Transition Setup. $\mathcal{G}_{\texttt{learn}}$ consists of $(M/4)$ MDPs , $\mathcal{M}_1, ..., \mathcal{M}_{M/4}$, which form the hard-to-learn example from Kwon et al. (2021) when no prospective side information is provided. In any of these MDPs in $\mathcal{G}_{\texttt{learn}}$, at the beginning, if an action a^*_{explore} is executed, the environment transitions to the starting of the hard-instance chain s^{hard}_1 with probability $\epsilon > 0$, or transitions to the starting of reference chain s^{ref}_1 with probability $1-\epsilon$. If any other action in $\mathcal{A}_{\text{explore}}$ is executed, the environment transitions to s^{ref}_1 with probability 1. For all other actions executed, the MDPs transition to a terminate-state s_{ter}.

$\mathcal{G}_{\texttt{ref}}$ consists of another $(M/4)$ MDPs, $\mathcal{M}_{M/4+1}, ..., \mathcal{M}_{M/2}$, which suppose to confuse the learning process in $\mathcal{G}_{\texttt{learn}}$. In these environments, dynamics in hard-to-learn chain and the reference chain are the same. Instead, at the beginning of an episode, if a^*_{explore} is executed, an MDP transitions to the starting of hard-to-learn chain s^{hard}_1 with probability $1-\epsilon$, or transitions to the starting of reference chain s^{ref}_1 with probability ϵ. If any other action in $\mathcal{A}_{\text{explore}}$ is executed, the environment transitions to s^{hard}_1 with probability 1. Like the group $\mathcal{G}_{\texttt{learn}}$, for all other actions executed, the MDPs in $\mathcal{G}_{\texttt{ref}}$ transition to s_{ter}.

The of the MDPs, indexed by $\mathcal{M}_{M/2+1}, ..., \mathcal{M}_M$, belong to the almost observable group $\mathcal{G}_{\texttt{obs}}$. In any of these MDPs in $\mathcal{G}_{\texttt{obs}}$, the environment always transitions to an absorbing state s_{ter} after playing an initial action. In each environment of this group $\mathcal{M}_m \in \mathcal{G}_{\texttt{obs}}$ where $m = M/2+1, ..., M$, playing a^m_{exploit} results with a reward of 1, and with 0 when playing any action different than a^m_{exploit}.

Prospective Side Information Setup. The prospective side information is a finite alphabet belongs and belongs to one of the $M+1$ disjoint sets $\mathcal{I}_1, \mathcal{I}_2, ..., \mathcal{I}_M, \mathcal{I}_{M+1}$. We let $\mathcal{I}_{M+1} := \{\iota_{\text{hard}}\}$ contains a single element, and all other disjoint sets have equal cardinality $|\mathcal{I}_1| = |\mathcal{I}_2| = ... = |\mathcal{I}_M| := |\mathcal{I}|$. For $\mathcal{M}_{M/2+1}, ..., \mathcal{M}_M$ in $\mathcal{G}_{\texttt{obs}}$, for each $m \in [M/2+1, M]$, the emission probability is give by $\mathbb{P}(\iota|m) = 1/(2|\mathcal{I}|)$ if $\iota \in \mathcal{I}_{m-M/2} \cup \mathcal{I}_m$, and 0 otherwise.

For all environments $\mathcal{M}_1, ..., \mathcal{M}_{M/4} \in \mathcal{G}_{\texttt{learn}}$ and $\mathcal{M}_{M/4+1}, ..., \mathcal{M}_{M/2} \in \mathcal{G}_{\texttt{ref}}$, for all $m \in [M/2]$, $\mathbb{P}(\iota_{\text{hard}}|m) = 1/2$, and $\mathbb{P}(\iota|m) = 0$ if $\iota \in \bigcup_{i=M/2+1}^{M} \mathcal{I}_i$. For all $\iota \in \bigcup_{i=1}^{M/2} \mathcal{I}_i$, we assign the probability of prospective side information $\mathbb{P}(\iota|m) \propto \frac{1+\alpha\varepsilon_{\iota,m}}{M|\mathcal{I}_m|}$, where each $\varepsilon_{\iota,m} \in \{-1, 1\}$ is decided in the following lemma:

Lemma C.1. *There exists a set of $\{\varepsilon_{\iota,m}\}_{\iota,m}$ such that for all $x \in \mathbb{R}^M$ it holds that $\|x\|_1 = 1$, $\|\mathbb{I}x\|_1 \geq \alpha' = \frac{\alpha}{128\sqrt{M}}$.*

Construction of Hard-to-Learn Chain for $\mathcal{G}_{\texttt{learn}}$, where $\mathcal{M}_1, ..., \mathcal{M}_d$ with $d = M/4$. This set is also depicted in the top part of Figure 1.

- At $t = 1$, *i.e.,* s^{hard}_1, there are three state-transition possibilities:

- $\mathcal{M}_1$: For all actions $a \in \mathcal{A}_{\text{control}}$ except a_1^*, we go to s_{ter}. For the action a_1^*, we go to s_2^{hard}.
- $\mathcal{M}_d$: For all actions $a \in \mathcal{A}_{\text{control}}$ except a_1^*, we go to s_2^{hard}. For the action a_1^*, we go to s_{ter}.
- $\mathcal{M}_2$, ..., $\mathcal{M}_{d-1}$: For all actions $a \in \mathcal{A}_{\text{control}}$, we go to s_2^{hard}.

- At time step $t = 2$, we again have three cases but now $\mathcal{M}_1$ and $\mathcal{M}_d$ would look the same:
  - $\mathcal{M}_1, \mathcal{M}_d$: For all actions $a \in \mathcal{A}_{\text{control}}$ except a_2^*, we go to s_{ter}. For the action a_2^*, we go to s_3^{hard}.
  - $\mathcal{M}_{d-1}$: For all actions $a \in \mathcal{A}_{\text{control}}$ except a_2^*, we go to s_3^{hard}. For the action a_2^*, we go to s_{ter}.
  - $\mathcal{M}_2$, ..., $\mathcal{M}_{d-2}$: For all actions $a \in \mathcal{A}_{\text{control}}$, we go to s_3^{hard}.

  ...

d. At time step $t = d$, we always transitionto s_{ter}, and there are two possibilities of getting rewards:
  - $\mathcal{M}_1$: For the action $a_d^* \in \mathcal{A}_{\text{control}}$, we get reward 1. For all other actions, we get rewards from $\text{Ber}(1/8)$.
  - $\mathcal{M}_2, ..., \mathcal{M}_d$: For all actions $a \in \mathcal{A}_{\text{control}}$, we get rewards from $\text{Ber}(1/8)$.

C.1. Proof of Lemma C.1

Proof. This can be shown by probabilistic arguments. Note that prospective side information that belongs to $\bigcup_{i=1}^{M/2} \mathcal{I}_i$ uniquely identifies the environment from $\mathcal{G}_{\text{no}}$, and thus

$$\begin{aligned}\|\mathbb{I}x\|_1 &= \|\mathbb{I}_{1:M/2}x\|_1 + \|\mathbb{I}_{M/2+1:M+1}x\|_1 \\ &\geq \frac{1}{2}\|x_{M/2+1:M}\|_1 + \max\left(0, \|\mathbb{I}_{M/2+1:M+1}x_{1:M/2}\|_1 - \frac{1}{2}\|x_{M/2+1:M}\|_1\right),\end{aligned}$$

where with slight abuse in notation, we denote $\mathbb{I}_{i:j}$ as the sub-matrix whose rows only correspond to one of prospective side information groups $\mathcal{I}_i, \mathcal{I}_{i+1}, ..., \mathcal{I}_j$. It is easy to check that if $\|\mathbb{I}_{M/2+1:M+1}x_{1:M/2}\|_1 \geq \frac{\alpha}{64\sqrt{M}}\|x_{1:M/2}\|_1$, then

$$\begin{aligned}\|\mathbb{I}x\|_1 &\geq \frac{1}{2}\|x_{M/2+1:M}\|_1 + \max\left(0, \frac{\alpha}{\sqrt{M}}\|x_{1:M/2}\|_1 - \frac{1}{2}\|x_{M/2+1:M}\|_1\right) \\ &\geq \frac{1}{2}\|x_{M/2+1:M}\|_1 + \max\left(0, \frac{\alpha}{\sqrt{M}} - \|x_{M/2+1:M}\|_1\right) \\ &\geq \frac{\alpha}{128\sqrt{M}}.\end{aligned}$$

Thus, it is sufficient to show that there exists $\{\varepsilon_{\iota,m}\}_{\iota,m}$ such that

$$\|\mathbb{I}_{M/2+1:M+1}x_{1:M/2}\|_1 \geq \|\mathbb{I}_{M/2+1:M}x_{1:M/2}\| \geq \frac{\alpha}{64\sqrt{M}}\|x_{1:M/2}\|_1.$$

Probabilistic Assignment. We set each $\varepsilon_{\iota,m}$ by an independent uniform sampling over $\{-1, 1\}$. We assume that $|\mathcal{I}_m|$ is sufficiently large, so that $\sum_{\iota\in\mathcal{I}} \varepsilon_{\iota,m}$ concentrates around 0 within $1/\sqrt{|\mathcal{I}|}$ and $1/\sqrt{|\mathcal{I}|}$ is sufficiently small.

Probabilistic Existence. To simplify the notation, we let $\mathbb{J} = \|\mathbb{I}_{M/2+1:M}x_{1:M/2}\|$ and $v = x_{1:M/2}$. Consider an $\gamma = \frac{\alpha}{256\sqrt{M}}$-cover, $\mathbb{B}_\gamma$ for the set $\{v \in \mathbb{R}^{M/2} : \|v\|_1 = 1\}$. Note that for each row of $\mathbb{J}$ and each $v \in \mathbb{B}_\gamma$,

$$|\mathbb{J}_\iota^\top v| = \frac{1}{M|\mathcal{I}|}\left|\sum_{m\in[M/2]} v_m + \alpha \cdot \sum_{m\in[M/2]} v_m\varepsilon_{\iota,m}\right|.$$

Without loss of generality, we assume $\sum_{m\in[M/2]} v_m \geq 0$. Note that the statistics of $W := |\sum_{m\in[M/2]} v_m\varepsilon_{\iota,m}|$, by Paley–Zygmund inequality (Paley & Zygmund, 1930),

$$\mathbb{P}\left(W \geq \frac{1}{2}\|v\|_2\right) \geq \frac{3}{16},$$

and thus with probability at least $3/32$, we have

$$\sum_{m\in[M/2]} v_m \varepsilon_{\iota,m} \geq \frac{1}{2\sqrt{M}} \implies |\mathbb{J}_\iota^\top v| \geq \frac{\alpha}{2\sqrt{M}}.$$

Since this holds for each row, and all $\epsilon_{\iota,m}$ are independent across the rows, at least $\frac{3}{64}(M/2)|\mathcal{I}_1|$ rows satisfies the above with probability at least $1-\exp(-(M/8)|\mathcal{I}_1|)$ from the concentration of the sum of independent Bernoulli random variables, which translates to

$$\|\mathbb{J}^\top v\|_1 \geq \frac{3\alpha}{128\sqrt{M}},$$

with probability $1-\exp(-(M/8)|\mathcal{I}_1|)$. Therefore, taking a union bound over $\mathbb{B}_\gamma$, we have

$$\|\mathbb{J}^\top v\|_1 \geq \frac{3\alpha}{128\sqrt{M}},$$

with probability $1-|\mathbb{B}_\gamma|\exp(-(M/8)|\mathcal{I}_1|) \geq 1-\exp(c_1 M\log(\gamma)-c_2|\mathcal{I}|)$ with proper absolute constants $c_1, c_2 > 0$. Then for arbitrary $v : \|v\|_1 = 1$, we can always find v_γ in $\mathbb{B}_\gamma$ such that $\|v - v_\gamma\| \leq \gamma$, and therefore

$$\|\mathbb{J}^\top v\|_1 \geq \|\mathbb{J}^\top v_\gamma\|_1 - \|\mathbb{J}^\top (v-v_\gamma)\|_1 \geq \frac{3\alpha}{128\sqrt{M}} - M\gamma.$$

Thus, setting $\gamma = o(\alpha/\sqrt{M})$ sufficiently small, for all $v : \|v\|_1 = 1$,

$$\|\mathbb{J}^\top v\|_1 \geq \frac{\alpha}{64\sqrt{M}}.$$

Since this probabilistic argument implies the existence of $\{\epsilon_{\iota,m}\}$, the proof is done. □

C.2. Proof of Lemma 5.2

This comes from the fundamental equality for sequential decision making information gain (see *e.g.,* Cesa-Bianchi & Lugosi (2006); Garivier et al. (2019); Kwon et al. (2023)). For completeness, we prove this. We can start from

$$\mathrm{KL}\left(\mathbb{P}_{\theta_0}^{\psi}(\tau^{1:K}), \mathbb{P}_{\theta}^{\psi}(\tau^{1:K})\right) = \mathbb{E}_{\theta_0}\left[\log\left(\frac{\mathbb{P}_{\theta_0}^{\psi}(\tau^{1:K-1})}{\mathbb{P}_{\theta}^{\psi}(\tau^{1:K-1})}\right)\right] + \mathbb{E}_{\theta_0}\left[\log\left(\frac{\mathbb{P}_{\theta_0}^{\psi}(\tau^{K}|\tau^{1:K-1})}{\mathbb{P}_{\theta}^{\psi}(\tau^{K}|\tau^{1:K-1})}\right)\right].$$

Note that in all models in our construction set $\Theta_{\text{hard}} \cup \{\theta_0\}$, $\mathbb{P}(\iota)$ and $\psi(a_t^k|\text{all histories until } k^{th} \text{ episode}, t^{th} \text{ step})$ are the same. Therefore, we have that

$$\begin{aligned}
&\mathbb{E}_{\theta_0}\left[\log\left(\frac{\mathbb{P}_{\theta_0}^{\psi}(\tau^{K}|\tau^{1:K-1})}{\mathbb{P}_{\theta}^{\psi}(\tau^{K}|\tau^{1:K-1})}\right)\right] \\
&= \mathbb{E}_{\theta_0}^{\psi}\left[\mathbb{E}_{\theta_0}^{\psi}\left[\sum_{\iota,a,a_{1:d}} \log\left(\frac{\mathbb{P}_{\theta_0}^{\psi}(\cdot|\iota,a,a_{1:d})}{\mathbb{P}_{\theta}^{\psi}(\cdot|\iota,a,a_{1:d})}\right)\mathbb{1}\left\{(\iota,a,a_{1:d})^K = (\iota,a,a_{1:d})\right\}\Big|\tau^{1:K-1}\right]\right] \\
&= \sum_{\iota,a,a_{1:d}} \mathbb{E}_{\theta_0}^{\psi}\left[\mathbb{E}_{\theta_0}^{\psi}\left[\log\left(\frac{\mathbb{P}_{\theta_0}^{\psi}(\cdot|\iota,a,a_{1:d})}{\mathbb{P}_{\theta}^{\psi}(\cdot|\iota,a,a_{1:d})}\right)\Big|\iota,a,a_{1:d}\right]\mathbb{1}\left\{(\iota,a,a_{1:d})^K = (\iota,a,a_{1:d})\right\}\right] \\
&= \sum_{\iota,a,a_{1:d}} \mathrm{KL}\left(\mathbb{P}_{\theta_0}(\cdot|\iota,a,a_{1:d}), \mathbb{P}_{\theta}(\cdot|\iota,a,a_{1:d})\right)\cdot\mathbb{E}_{\theta_0}^{\psi}\left[\mathbb{1}\left\{(\iota,a,a_{1:d})^K = (\iota,a,a_{1:d})\right\}\right],
\end{aligned}$$

where the second equality is an application of the tower rule, and the last equality is due to the choice of action purely depends on the history and exploration strategy ψ, and does not depend on underlying models. Applying this recursively in K, and denoting $N^a_{\psi,\iota,a_{1:d}}(K)$ as the number of times action a was executed at the initial step, $a_{1:d}$ in the next d steps under prospective side information ι. Thus, we have

$$\mathrm{KL}\left(\mathbb{P}_{\theta_0}^{\psi}(\tau^{1:K}), \mathbb{P}_{\theta}^{\psi}(\tau^{1:K})\right) = \sum_{\iota,a,a_{1:d}} \mathbb{E}_{\theta_0}\left[N^a_{\psi,\iota,a_{1:d}}(K)\right]\cdot\mathrm{KL}\left(\mathbb{P}_{\theta_0}(\cdot|\iota,a,a_{1:d}), \mathbb{P}_{\theta}(\cdot|\iota,a,a_{1:d})\right),$$

Note that when playing $a \neq a^*_{\text{explore}}$, the hard instance and the reference model behave the same, yielding the result.

C.3. Proof of Lemma 5.3

We first check the following inequality:

$$\mathrm{KL}\left(\mathbb{P}_{\theta_0}(\cdot|\iota_{\text{hard}}, a^*_{\text{explore}}, a^*_{1:d}), \mathbb{P}_{\theta}(\cdot|\iota_{\text{hard}}, a^*_{\text{explore}}, a^*_{1:d})\right).$$

The point is that until seeing the last time-step event, the distribution of histories are the same in all environments. To see this, at the initial time step given the prospective side information ι_{hard}, the belief over latent contexts are all equal to $2/M$ for all MDPs in $\mathcal{G}_{\texttt{learn}}$ and $\mathcal{G}_{\texttt{ref}}$. Thus, the probability of transitioning to s_1^{hard} is $1/2$ by executing a^*_{explore} (if the environment transitions to s_1^{ref}, or any other action is executed, then the future distribution on of all events are exactly the same in all hard and reference instances). In the middle of the hard-instance chain, at s_t^{hard}, the probability of moving to the next state conditioned on the past is $1 - 1/(d-t+1)$. However, the true posterior probability over MDPs from $\mathcal{G}_{\texttt{ref}}$ at this point is given by:

$$\mathbb{P}(m|\iota_{\text{hard}}, a_{1:t}, s_t^{\text{hard}}) = \epsilon/(d-t+1),$$

for all $m = 1, 2, ..., M/4$ with non-zero posteriors (since we eliminated MDPs from the set after gathering information in a certain way). On the other hand,

$$\mathbb{P}(m|\iota_{\text{hard}}, a_{1:t}, s_t^{\text{hard}}) = 4(1-\epsilon)/M,$$

for all $m = M/4+1, ..., M/2$, *i.e.,* MDPs from $\mathcal{G}_{\texttt{ref}}$. Thus, at the last time step, the chance of observing the reward 1 conditioned on the history that we reached s_d^{hard} with the optimal action sequence $a^*_{1:d}$, is $1/8 + O(\epsilon)$ in hard instances, and $1/8$ in the reference model. Thus, the KL divergence between the two models takes the following form:

$$\begin{aligned}
&\mathrm{KL}\left(\mathbb{P}_{\theta_0}(\cdot|\iota_{\text{hard}}, a^*_{\text{explore}}, a^*_{1:d}), \mathbb{P}_{\theta}(\cdot|\iota_{\text{hard}}, a^*_{\text{explore}}, a^*_{1:d})\right) \\
&= \sum_{r_d \in \{0,1\}} \mathbb{P}_{\theta_0}(r_d, s^{\text{hard}}_{1:d}|\iota_{\text{hard}}, a^*_{\text{explore}}, a^*_{1:d}) \cdot \log\left(\frac{\mathbb{P}_{\theta_0}(r_d, s^{\text{hard}}_{1:d}|\iota_{\text{hard}}, a^*_{\text{explore}}, a^*_{1:d})}{\mathbb{P}_{\theta}(r_d, s^{\text{hard}}_{1:d}|\iota_{\text{hard}}, a^*_{\text{explore}}, a^*_{1:d})}\right) \\
&= \mathbb{P}_{\theta_0}(s^{\text{hard}}_{1:d}|\iota_{\text{hard}}, a^*_{\text{explore}}, a^*_{1:d}) \sum_{r_d \in \{0,1\}} \mathbb{P}_{\theta_0}(r_d|s^{\text{hard}}_{1:d}, \iota_{\text{hard}}, a^*_{\text{explore}}, a^*_{1:d}) \log\left(\frac{\mathbb{P}_{\theta_0}(r_d|s^{\text{hard}}_{1:d}, \iota_{\text{hard}}, a^*_{\text{explore}}, a^*_{1:d})}{\mathbb{P}_{\theta}(r_d|s^{\text{hard}}_{1:d}, \iota_{\text{hard}}, a^*_{\text{explore}}, a^*_{1:d})}\right) \\
&\leq \frac{1}{2d} \cdot \mathrm{KL}(\mathrm{Ber}(1/8), \mathrm{Ber}(1/8 + O(\epsilon)) \lesssim \epsilon^2/M.
\end{aligned}$$

For other inequalities, note that for any trajectory with any $a \neq a^*_{\text{explore}}$, for all ι and $a_{1:d} \in \mathcal{A}^{\otimes d}$, the marginal distribution is always the same in all hard-instances and the reference model. The marginal distribution is also the same when transitioning to s_1^{ref} even if a^*_{explore} is executed at the initial time. Thus, we can focus on the case when the action at the initial time step is a^*_{explore}, and the environment transitions to s_1^{hard}. If this is the case, for all $s_{2:d}$,

$$\begin{aligned}
\mathbb{P}((s_1^{\text{hard}}, s_{2:d}), r_d|\iota, a^*_{\text{explore}}, a_{1:d}) &= \sum_{m \in [M/2]} p_m(\iota) \mathbb{P}((s_1^{\text{hard}}, s_{2:d}), r_d|a^*_{\text{explore}}, a_{1:d}, m) \\
&= \sum_{m \in [M/2]} p_m(\iota) \mathbb{P}(s_1^{\text{hard}}|a^*_{\text{explore}}, m) \mathbb{P}(s_{2:d}, r_d|s_1^{\text{hard}}, a_{1:d}, m).
\end{aligned}$$

Note that

$$p_m(\iota) = \frac{\mathbb{P}(\iota|m)}{\sum_{m'} \mathbb{P}(\iota|m')},$$

and in all models, and since for all $m \in [M/2]$,

$$\mathbb{P}(\iota|m) \propto (1 + \alpha\varepsilon_{\iota,m}),$$

we can observe that

$$p_m(\iota) = \frac{(1+\alpha\varepsilon_{i,m})}{M/2 + \sum_{m' \in [M/2]}(1+\alpha\varepsilon_{\iota,m'})} = \frac{(1+O(\alpha))}{M},$$

Therefore, in all instances,

$$\begin{aligned}\mathbb{P}(s_1^{\text{hard}}, s_{2:d}, r_d|\iota, a^*_{\text{explore}}, a_{1:d}) &= \epsilon \sum_{m\in[M/4]} p_m(\iota)\mathbb{P}(s_{2:d}, r_d|s_1^{\text{hard}}, a_{1:d}, m) \\ &\quad + (1-\epsilon) \sum_{m\in[M/4+1,M/2]} p_m(\iota)\mathbb{P}(s_{2:d}, r_d|s_1^{\text{hard}}, a_{1:d}, m) \\ &= \frac{(1+O(\alpha))\epsilon}{4} \sum_{m\in[M/4]} \frac{4}{M}\mathbb{P}(s_{2:d}, r_d|s_1^{\text{hard}}, a_{1:d}, m) \\ &\quad + (1-\epsilon) \sum_{m\in[M/4+1,M/2]} p_m(\iota)\mathbb{P}(s_{2:d}, r_d|s_1^{\text{hard}}, a_{1:d}, m).\end{aligned}$$

Now comparing this probability between any hard-instance $\theta \in \Theta_{\text{hard}}$ and reference model, note that

$$\mathbb{P}_{\theta_0}(s_{2:d}, r_d|s_1^{\text{hard}}, a_{1:d}, m) = \mathbb{P}_{\theta}(s_{2:d}, r_d|s_1^{\text{hard}}, a_{1:d}, m),$$

for all $m \in [M/4+1, M/2]$, and

$$\sum_{m\in[M/4]} \frac{4}{M}\mathbb{P}_{\theta_0}(s_{2:d}, r_d|s_1^{\text{hard}}, a_{1:d}, m) = \sum_{m\in[M/4]} \frac{4}{M}\mathbb{P}_{\theta}(s_{2:d}, r_d|s_1^{\text{hard}}, a_{1:d}, m),$$

for all $s_{2:d} \neq s_{2:d}^{\text{hard}}$ or $a_{1:d} \neq a^*_{1:d}$, and

$$\begin{aligned}\sum_{m\in[M/4]} \frac{4}{M}\mathbb{P}_{\theta_0}(s_{2:d}^{\text{hard}}, r_d = 1|s_1^{\text{hard}}, a^*_{1:d}, m) &= \frac{4}{M}\cdot\frac{1}{8} = \frac{1}{2M}, \\ \sum_{m\in[M/4]} \frac{4}{M}\mathbb{P}_{\theta}(s_{2:d}^{\text{hard}}, r_d|s_1^{\text{hard}}, a^*_{1:d}, m) &= \frac{4\cdot\mathbb{1}\{r_d = 1\}}{M}.\end{aligned}$$

Therefore,

$$\begin{aligned}&|\mathbb{P}_{\theta}(s_1^{\text{hard}}, s_{2:d}, r_d|\iota, a^*_{\text{explore}}, a_{1:d}) - \mathbb{P}_{\theta_0}(s_1^{\text{hard}}, s_{2:d}, r_d|\iota, a^*_{\text{explore}}, a_{1:d})| \\ &= O(\alpha\epsilon) \sum_{m\in[M/4]} \frac{4}{M}\mathbb{P}_{\theta_0}(s_{2:d}, r_d|s_1^{\text{hard}}, a_{1:d}, m) + O(\epsilon)\frac{\mathbb{1}\left\{s_{2:d} = s_{2:d}^{\text{hard}}, a_{1:d} = a^*_{1:d}\right\}}{M},\end{aligned}$$

and also we note that

$$\begin{aligned}&\sum_{m\in[M/4]} \frac{4}{M}\mathbb{P}_{\theta_0}(s_{2:d}, r_d|s_1^{\text{hard}}, a_{1:d}, m) = O(\mathbb{P}_{\theta_0}(s_1^{\text{hard}}, s_{2:d}, r_d|\iota, a^*_{\text{explore}}, a_{1:d})), \\ &\mathbb{P}_{\theta_0}(s_{1:d}^{\text{hard}}, r_d|\iota, a^*_{\text{explore}}, a^*_{1:d}) = O(1/M).\end{aligned}$$

Therefore, we can ensure that

$$\begin{aligned}&|\mathbb{P}_{\theta}(s_1^{\text{hard}}, s_{2:d}, r_d|\iota, a^*_{\text{explore}}, a_{1:d}) - \mathbb{P}_{\theta_0}(s_1^{\text{hard}}, s_{2:d}, r_d|\iota, a^*_{\text{explore}}, a_{1:d})| \\ &\quad \le O(\alpha\epsilon)\mathbb{P}_{\theta_0}(s_1^{\text{hard}}, s_{2:d}, r_d|\iota, a^*_{\text{explore}}, a_{1:d}), \\ &\sum_{s_{2:d}, r_d} |\mathbb{P}_{\theta}(s_1^{\text{hard}}, s_{2:d}, r_d|\iota, a^*_{\text{explore}}, a_{1:d}) - \mathbb{P}_{\theta_0}(s_1^{\text{hard}}, s_{2:d}, r_d|\iota, a^*_{\text{explore}}, a_{1:d})| \le O(\alpha\epsilon),\end{aligned}$$

for all $a_{1:d} \neq a^*_{1:d}$, and similarly for $a^*_{1:d}$,

$$\begin{aligned}&|\mathbb{P}_{\theta}(s_1^{\text{hard}}, s_{2:d}, r_d|\iota, a^*_{\text{explore}}, a^*_{1:d}) - \mathbb{P}_{\theta_0}(s_1^{\text{hard}}, s_{2:d}, r_d|\iota, a^*_{\text{explore}}, a^*_{1:d})| \\ &\quad \le O(\epsilon)\mathbb{P}_{\theta_0}(s_1^{\text{hard}}, s_{2:d}, r_d|\iota, a^*_{\text{explore}}, a^*_{1:d}), \\ &\sum_{s_{2:d}, r_d} |\mathbb{P}_{\theta}(s_1^{\text{hard}}, s_{2:d}, r_d|\iota, a^*_{\text{explore}}, a^*_{1:d}) - \mathbb{P}_{\theta_0}(s_1^{\text{hard}}, s_{2:d}, r_d|\iota, a^*_{\text{explore}}, a^*_{1:d})| \le O(\epsilon).\end{aligned}$$

Finally, to bound the KL-divergence, using $\log(x) \le x - 1$,

$$\begin{aligned}
&\mathrm{KL}\left(\mathbb{P}_{\theta_0}(\cdot|\iota, a^*_{\text{explore}}, a_{1:d}), \mathbb{P}_{\theta}(\cdot|\iota, a^*_{\text{explore}}, a_{1:d})\right) \\
&= \sum_{s_{1:d}, r_d} \mathbb{P}_{\theta_0}(r_d, s_{1:d}|\iota, a^*_{\text{explore}}, a_{1:d}) \cdot \log\left(\frac{\mathbb{P}_{\theta_0}(r_d, s_{1:d}|\iota, a^*_{\text{explore}}, a_{1:d})}{\mathbb{P}_{\theta}(r_d, s_{1:d}|\iota, a^*_{\text{explore}}, a_{1:d})}\right) \\
&\le \sum_{s_{1:d}, r_d} \mathbb{P}_{\theta_0}(r_d, s_{1:d}|\iota, a^*_{\text{explore}}, a_{1:d}) \cdot \left(\frac{\mathbb{P}_{\theta_0}(r_d, s_{1:d}|\iota, a^*_{\text{explore}}, a_{1:d})}{\mathbb{P}_{\theta}(r_d, s_{1:d}|\iota, a^*_{\text{explore}}, a_{1:d})} - 1\right) \\
&= \sum_{s_{1:d}, r_d} \frac{\left|\mathbb{P}_{\theta_0}(r_d, s_{1:d}|\iota, a^*_{\text{explore}}, a_{1:d}) - \mathbb{P}_{\theta}(r_d, s_{1:d}|\iota, a^*_{\text{explore}}, a_{1:d})\right|^2}{\mathbb{P}_0(r_d, s_{1:d}|\iota, a^*_{\text{explore}}, a_{1:d})} \\
&= \sum_{s_{2:d}, r_d} \frac{\left|\mathbb{P}_{\theta_0}(r_d, s^{\text{hard}}_1, s_{2:d}|\iota, a^*_{\text{explore}}, a_{1:d}) - \mathbb{P}_{\theta}(r_d, s^{\text{hard}}_1, s_{2:d}|\iota, a^*_{\text{explore}}, a_{1:d})\right|^2}{\mathbb{P}_{\theta_0}(r_d, s^{\text{hard}}_1, s_{2:d}|\iota, a^*_{\text{explore}}, a_{1:d})},
\end{aligned}$$

which is $O(\alpha\epsilon)^2$ if $a_{1:d} \neq a^*_{1:d}$, and $O(\epsilon^2)$ if $a_{1:d} = a^*_{1:d}$. This concludes the proof of the lemma.

C.4. Proof of Theorem 4.2

Proof. Suppose any learning strategy (algorithm). Note that an $\epsilon/4$-optimal policy for any given $\theta \in \Theta_{\text{hard}}$ should be able to play the correct action sequence $a^*_{1:d}$ of θ whenever the prospective side information is ι_{hard}. On the other hand, by pigeon hole principle, for any algorithm ψ with any choice of K, there must exist at least one action-sequence $a^*_{1:d}$ and a^*_{explore} such that,

$$\begin{aligned}
\sum_{\iota} \mathbb{E}_0[N^{\text{explore}}_{\psi,\iota,a^*_{1:d}}(K)] &= \min_{a \in \mathcal{A}_{\text{explore}}, a_{1:d}} \left(\sum_{\iota} \mathbb{E}_0[N^{a}_{\psi,\iota,a_{1:d}}(K)]\right) \le |\mathcal{A}_{\text{control}}|^{-(d+1)} \cdot K, \\
\sum_{\iota \neq \iota_{\text{hard}}, a_{1:d}} \mathbb{E}_0[N^{\text{explore}}_{\psi,\iota,a_{1:d}}(K)] &= \min_{a \in \mathcal{A}_{\text{explore}}} \left(\textstyle\sum_{\iota \neq \iota_{\text{hard}}, a_{1:d}} \mathbb{E}_0[N^{a}_{\psi,\iota,a_{1:d}}(K)]\right) \le |\mathcal{A}_{\text{explore}}|^{-1} \cdot K.
\end{aligned}$$

Let K_0 be the largest number such that with this choice of $a^*_{1:d}$ and a^*_{explore}, equation (9) does not hold, i.e.,

$$\mathbb{E}_{\theta_0}[N^{\text{explore}}_{\psi,\iota_{\text{hard}},a^*_{1:d}}(K_0+1)] \gtrsim \frac{1}{\epsilon^2}, \text{ or } \sum_{\iota \neq \iota_{\text{hard}}, a_{1:d}} \mathbb{E}_{\theta_0}[N^{\text{explore}}_{\psi,\iota,a_{1:d}}(K_0+1)] \gtrsim \frac{1}{\alpha^2\epsilon^2}.$$

We also note that

$$N^{a}_{\psi,\iota,a_{1:d}}(K') \ge N^{a}_{\psi,\iota,a_{1:d}}(K), \tag{15}$$

for any $K' > K$ with probability 1.

Note that if we play $a \notin \mathcal{A}_{\text{exploit}}$ whenever $\iota \neq \iota_{\text{hard}}$, we incur at least $(1/8)$-regret. On the other hand, if we do not play a^*_{explore} or $a_{1:d} \neq a^*_{1:d}$ when $\iota = \iota_{\text{hard}}$, we incur at least $\epsilon/(2M)$-regret. Thus, the total regret of the algorithm in the hard-instance $\theta \in \Theta_{\text{hard}}$ is given by

$$\mathrm{Regret}_{\theta}(K) \ge \sum_{a \in \mathcal{A}, a_{1:d} \neq a^*_{1:d}} \mathbb{E}_{\theta}[N^{a}_{\psi,\iota_{\text{hard}},a_{1:d}}(K)] \cdot \frac{\epsilon}{2M} + \sum_{\iota \neq \iota_{\text{hard}}, a \in \mathcal{A}_{\text{explore}}, a_{1:d}} \mathbb{E}_{\theta}[N^{a}_{\psi,\iota,a_{1:d}}(K)] \cdot \frac{1}{8}.$$

On the other hand, the regret in the reference model satisfies

$$\mathrm{Regret}_{0}(K) \ge \sum_{\iota \neq \iota_{\text{hard}}, a \in \mathcal{A}_{\text{explore}}, a_{1:d}} \mathbb{E}_{0}[N^{a}_{\psi,\iota,a_{1:d}}(K)] \cdot \frac{1}{8}.$$

Now we consider three cases:

Case (1). If $(A/3)^{-d}K_0 \le K \le K_0$, then the condition in equation (9) cannot be satisfied and therefore, (with proper absolute constants) $\mathrm{KL}\left(\mathbb{P}_0^{\psi}(\tau^{1:K}), \mathbb{P}_{\text{hard}}^{\psi}(\tau^{1:K})\right) \le 1/128$, implying $d_{\mathrm{TV}}\left(\mathbb{P}_0^{\psi}(\tau^{1:K}), \mathbb{P}_{\text{hard}}^{\psi}(\tau^{1:K})\right) \le 1/16$ by Pinsker's inequality. Note that

$$\sum_{a\in\mathcal{A}, a_{1:d}\neq a^*_{1:d}} \mathbb{E}_0[N^a_{\psi,\iota_{\text{hard}},a_{1:d}}(K)] = \frac{K}{2} - \mathbb{E}_0[N^{\text{explore}}_{\psi,\iota_{\text{hard}},a^*_{1:d}}(K)] \ge \frac{1}{3}K,$$

and thus, since the sum is always bounded by K, with probability at least $1/6$,

$$\sum_{a\in\mathcal{A}, a_{1:d}\neq a^*_{1:d}} N^a_{\psi,\iota_{\text{hard}},a_{1:d}}(K) \ge K/6,$$

in the reference model. Here, we used the fact that for any non-negative random variable A that is almost surely bounded by K, for all $0 \le x \le K$,

$$\mathbb{E}[A] = \mathbb{E}[A|A>x]\cdot\mathbb{P}(A>x) + \mathbb{E}[A|A\le x]\cdot\mathbb{P}(A\le x) \le K\mathbb{P}(A>x) + x.$$

Therefore in the hard instance $\mathcal{M}$:

$$\sum_{a\in\mathcal{A}, a_{1:d}\neq a^*_{1:d}} N^a_{\psi,\iota_{\text{hard}},a_{1:d}}(K) \ge K/6,$$

with probability at least $1/16$, confirming $\sum_{a\in\mathcal{A}, a_{1:d}\neq a^*_{1:d}} \mathbb{E}[N^a_{\psi,\iota_{\text{hard}},a_{1:d}}(K)] \ge K/256$. Thus in this case,

$$\text{Regret}_\theta(K) \gtrsim \frac{K\epsilon}{M}.$$

Case (2). Suppose $K > K_0$ and $\mathbb{E}_0[N^{\text{explore}}_{\psi,\iota_{\text{hard}},a^*_{1:d}}(K_0+1)] \gtrsim \frac{1}{\epsilon^2}$ but $\mathbb{E}_0[N^{\text{explore}}_{\psi,\iota_{\text{hard}},a^*_{1:d}}(K_0)] \lesssim \frac{1}{\epsilon^2}$. Note that by the same argument, we know that

$$\sum_{a\in\mathcal{A}, a_{1:d}\neq a^*_{1:d}} \mathbb{E}[N^a_{\psi,\iota_{\text{hard}},a_{1:d}}(K_0)] \ge K_0/256,$$

and since equation (15) holds with probability 1, we have

$$\sum_{a\in\mathcal{A}, a_{1:d}\neq a^*_{1:d}} \mathbb{E}[N^a_{\psi,\iota_{\text{hard}},a_{1:d}}(K)] \ge K_0/256.$$

On the other hand, to satisfy this condition, we need at least $K_0 \ge \frac{(A/3)^d}{\epsilon^2}$. Thus, plugging this to the regret bound, we have that

$$\text{Regret}_\theta(K) \ge \frac{K_0}{256}\frac{\epsilon}{2M} \gtrsim \frac{(A/3)^d}{M\epsilon}.$$

Case (3). Finally, suppose $K > K_0$ and $\mathbb{E}_0[N^{\text{explore}}_{\psi,\iota,a_{1:d}}(K_0+1)] \gtrsim \frac{1}{\alpha^2\epsilon^2}$ but $\mathbb{E}_0[N^{\text{explore}}_{\psi,\iota,a_{1:d}}(K_0)] \lesssim \frac{1}{\alpha^2\epsilon^2}$. Then by construction (due to the choice of a^*_{explore}), we have

$$\sum_{\iota\neq\iota_{\text{hard}}, a\in\mathcal{A}_{\text{explore}}, a_{1:d}} \mathbb{E}_0[N^a_{\psi,\iota,a_{1:d}}(K)] \ge \frac{(A/3)}{\alpha^2\epsilon^2},$$

and thus the regret incurred in the reference model is

$$\text{Regret}_0(K) \ge \frac{(A/3)}{\alpha^2\epsilon^2}\frac{1}{8} \gtrsim \frac{A}{\alpha^2\epsilon^2}.$$

Combining the three cases, for any algorithm ψ, we can conclude that

$$\max_{\theta\in\Theta_{\text{hard}}\cup\{\theta_0\}} \text{Regret}_\theta(K) \gtrsim \min\left(\frac{(A/3)^d}{M\epsilon}, \frac{A}{\alpha^2\epsilon^2}, K\epsilon\right).$$

□