Provably Efficient Representation Learning with Tractable Planning in Low-Rank POMDP

Jiacheng Guo [1 2] Zihao Li [1] Huazheng Wang [3] Mengdi Wang [1] Zhuoran Yang [4] Xuezhou Zhang [1]

Abstract

In this paper, we study representation learning in partially observable Markov Decision Processes (POMDPs), where the agent learns a decoder function that maps a series of high-dimensional raw observations to a compact representation and uses it for more efficient exploration and planning. We focus our attention on the sub-classes of γ-*observable* and *decodable POMDPs*, for which it has been shown that statistically tractable learning is possible, but there has not been any computationally efficient algorithm. We first present an algorithm for decodable POMDPs that combines maximum likelihood estimation (MLE) and optimism in the face of uncertainty (OFU) to perform representation learning and achieve efficient sample complexity, while only calling supervised learning computational oracles. We then show how to adapt this algorithm to also work in the broader class of γ-observable POMDPs.

1. Introduction

Markov Decision Processes (MDPs) are commonly used in reinforcement learning to model problems across a range of applications that involve sequential decision-making. However, MDPs assume that the agent has perfect knowledge of the current environmental state, which is often not realistic. To address this, Partially Observable Markov Decision Processes (POMDPs) have been introduced as an extension of MDPs (Cassandra, 1998; Murphy, 2000; Braziunas, 2003). In POMDPs, the agent does not have direct access to the environmental state. Instead, it receives observations that are sampled from a state-dependent distribution. POMDPs are an important class of models for decision-making that is increasingly being used to model complex real-world applications, ranging from robotics and navigation to healthcare, finance, manufacturing (Roy et al., 2006; Chen et al., 2016; Ghandali et al., 2018; Liu et al., 2020).

POMDPs differ from MDPs in the presence of *non-Markovian* observations. In the MDP setting, we assume that the system follows a Markovian law of transition induced by the agent's policy, meaning that the current state of the environment depends solely on the previous state and the action taken. As a result, there exists a Markovian optimal policy, i.e. a policy whose action only depends on the current state. However, with the assumption of unobserved hidden states and therefore the lack of Markovian property, generally, the optimal policy of POMDP depends on the full history. This causes great difficulties in both computational and statistical aspects, as the agent has to maintain a long-term *memory* while performing learning and planning. In fact, the agent can only learn the posterior distribution of the latent state given the whole history, known as the belief state, and then maps the belief state to the action. Therefore, learning POMDP is well-known to be intractable in general, even with a small state and observation space (Papadimitriou & Tsitsiklis, 1987a; Mundhenk et al., 2000; Golowich et al., 2022b).

Nevertheless, this doesn't rule out special problem structures that enable statistically efficient algorithms. In particular, Katt et al. (2018) and Liu et al. (2022a) study a setting called γ-*observable* POMDP, i.e. POMDP whose omission matrix is full-rank, and achieve polynomial sample complexities. Another more tractable class is called the *decodable* POMDP, in which we can decode the latent state by L-step back histories (Efroni et al., 2022b). Subsequent works on provably efficient learning in POMDP have since found broader classes of statistically tractable problem instances that generalized the above setting (Zhan et al., 2022; Uehara et al., 2022b; Liu et al., 2022b). One common drawback of the aforementioned works, however, is their computational intractability. Almost all of these algorithms follow the algorithmic template of *optimistic planning*, where the algorithm iteratively refines a version space of plausible functions in the function class. To achieve this, these meth-

[1]Department of Electrical and Computer Engineering, Princeton University, Princeton, NJ, USA [2]School of Mathematical Sciences, Fudan University, Shanghai, China [3]School of Electrical Engineering and Computer Science, Oregon State University, OR, USA [4]Department of Statistics and Data Science, Yale University, NH, USA. Correspondence to: Mengdi Wang <mengdiw@princeton.edu>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

ods perform an iteration and elimination procedure over all remaining functions in the function class in every iteration of the algorithm, making them unimplementable for large function classes. In this paper, we take a representation learning approach to overcome this issue.

Recent works have found that by performing representation learning and using the learned feature to construct exploration bonuses, one can achieve efficient reinforcement learning in rich-observation environments both in theory (Silver et al., 2018; Stooke et al., 2021; Uehara et al., 2021b) and in practice (Guo et al., 2022). However, learning representation is more subtle in POMDPs. The main challenge is two-folded: (1) To efficiently learn a near-optimal policy, the agent must be able to construct a representation of the history that compactly captures the belief state. This causes difficulty in computation, as the agent will suffer from the curse of *long memory* (Golowich et al., 2022a; Liu et al., 2022a). (2) Without direct knowledge of the underlying state, it is hard to design an exploration bonus that balances *exploration and exploitation*. Motivated by the above challenges, the main inquiry we aim to answer in this study is:

Can we design a representation-learning-aided RL algorithm for POMDPs that is both statistically efficient and amenable to efficient implementation?

In this work, we give an affirmative answer to the question. We consider the subclass of γ-*observabile* and *decodable* POMDP. To handle (1), we consider a short memory learning scheme that considers a fixed-size history instead of the whole history, which is common in POMDP planning (Golowich et al., 2022b). To handle (2), we follow the principle of *optimism in the face of uncertainty (OFU)*, which chooses actions according to the upper bound of the value function. Such a method is typical for balancing exploration and exploitation trade-offs in online MDP learning (Cai et al., 2020; Jin et al., 2020b; Yang & Wang, 2020; Ayoub et al., 2020a). We also remark that our UCBVI-type algorithm is computationally efficient, as it avoids intractable optimistic planning (e.g., UCRL (Auer et al., 2008)). Specifically, we present a representation learning-based reinforcement learning algorithm for solving POMDPs with low-rank latent transitions, which is common for representation learning in MDP cases (Agarwal et al., 2020; Uehara et al., 2021b; Zhan et al., 2022).

2. Related Work

Our work is built upon the bodies of literature on both (i) reinforcement learning in POMDPs and (ii) representation learning. In this section, we will focus on the related works in these two directions.

Learning in POMDPs Provably efficient RL methods for POMDPs have been studied in a number of recent works (Li et al., 2009; Guo et al., 2016; Katt et al., 2018; Jin et al., 2020a; Jafarnia-Jahromi et al., 2021; Liu et al., 2022a). Even when the underlying dynamics of the POMDP are known, simply planning is still hard and relies on *short-memory approximation* (Papadimitriou & Tsitsiklis, 1987b). Moreover, when learning POMDP, the estimation of the model is computationally hard (Mossel & Roch, 2005), and learning the POMDPs is statistically intractable (Krishnamurthy et al., 2016).

However, this doesn't rule out the possibility of finding an efficient algorithm for a particular class for POMDP. A line of work studies a wide sub-classes of POMDPs and has achieved positive results. Guo et al. (2016) and Azizzadenesheli et al. (2016) use spectral methods to learn POMDPs and obtain polynomial sample complexity results without addressing the strategic exploration. (Jafarnia-Jahromi et al., 2021) uses the posterior sampling technique to learn POMDPs in the Bayesian setting, with time, the posterior distribution will converge to the true distribution.

The *observability assumption*, or the *weakly-revealing* assumption, has been widely studied for learning in POMDPs. It assumes that the distribution on observations space can recover the distribution on the latent states. It is a very rich subset as it contains the tractable settings in (Guo et al., 2016; Jin et al., 2020b) such as the overcomplete setting. By incorporating observability and assuming a computation oracle such as optimistic planning, (Katt et al., 2018; Liu et al., 2022a) achieve favorable polynomial sample complexities. At the same time, Golowich et al. (2022b) proposed algorithms that can achieve quasi-polynomial sample and computational complexity.

Another common assumption used in POMDPs is the *decodability* assumption, which assumes we can reveal the latent state by L-step back histories (Efroni et al., 2022b). Efroni et al. (2022b) obtained polynomial sample complexities. It can be regarded as a generalization of block MDPs (Krishnamurthy et al., 2016), in which the latent state can be uniquely determined by the current observation. More generally, in an L-step *decodable* POMDP, the latent state can be uniquely decoded from the most recent history (of observations and actions) of a short length L. We remark on the existing literature in block MDPs or decodable POMDPs (e.g. Krishnamurthy et al. (2016); Du et al. (2019); Misra et al. (2020); Efroni et al. (2022b); Liu et al. (2022a))

Representation Learning in RL There has been considerable progress on provable representation learning for RL in recent literature (Ayoub et al., 2020b; Agarwal et al., 2020; Modi et al., 2021; Uehara et al., 2021a). Modi et al. (2021); Zhang et al. (2022b) present the model-free approach base

Algorithm	Setting	Sample Complexity	Computation
Liu et al. (2022a)	Tabular	$\mathrm{poly}(S, A, O, H, \gamma^{-1})/\epsilon^2$	Version Space
Golowich et al. (2022a)	Tabular	quasi-poly$(S, A, O, H, (OA)^{\gamma^{-1}}, \epsilon^{-1})$	quasi-poly$(S, A, O, H, (OA)^{\gamma^{-1}}, \epsilon^{-1})$
Uehara et al. (2022b)	Low-rank	quasi-poly$(H, d, A^{c/\gamma}, \log|\mathcal{M}|)/\epsilon^2$	Version Space
Zhan et al. (2022)	Low-rank	$\mathrm{poly}(H, d, A, 1/\gamma, \log|\mathcal{M}|, \log O)/\epsilon^2$	Version Space
Wang et al. (2022)	Low-rank*	quasi-poly$(H, d, A^{c/\gamma}, \log|\mathcal{M}|)/\epsilon^2$	Version Space
PORL^2 (ours)	Low-rank	quasi-poly$(H, d, A^{c/\gamma}, \log|\mathcal{M}|)/\epsilon^2$	MLE Oracle+LSVI-LLR

Table 1. Comparison to all existing works that solve γ-observable POMDP with formal sample complexity guarantees. Among them, Liu et al. (2022a) and Golowich et al. (2022a) study the tabular setting and their algorithm cannot easily incorporate function approximations. Golowich et al. (2022a) proposed the only algorithm for POMDP that is known to have a formal computational complexity guarantee, and yet the algorithm relies heavily on iterating over the observation space, which cannot be done when the observation space is large or even infinite. Among the algorithms that handle low-rank observable POMDPs, that of Zhan et al. (2022) escapes the exponential dependency on γ at a cost of an explicit dependency on the size of the observation space O. Computationally, all existing algorithms other than Golowich et al. (2022a) fall into the category of version-space learners, where the algorithm must keep track of the set remaining plausible functions in the function class and eliminate the ones that are inconsistent with existing observations. Such a procedure will have a computational complexity scale linearly with the size of the function class, which is generally considered to be inefficient, and particularly not amendable for modern neural-network-based implementation. In contrast, our algorithm only relies on calling an MLE computational oracle, which is standard in supervised learning and amendable to efficient implementation[2] and calling the LSVI-LLR algorithm, which has a $\mathrm{poly}(H, A, d)$ computational complexity.

on block MDPs and (Uehara et al., 2021b; Agarwal et al., 2020) study the model-based approach through MLE on general low-rank MDP setting. These works focus on fully observable settings, only consider an environment with a Markovian transition. For POMDPs, Wang et al. (2022) learn the representation with a constant past sufficiency assumption, and their sample complexity has an exponential dependence on that constant. Uehara et al. (2022a) uses an actor-critic style algorithm to capture the value link functions with the assumption that the value function is linear in historical trajectory, which can be too strong in practice. Zhan et al. (2022); Liu et al. (2022a) uses MLE to construct a confidence set for the model of POMDP or PSR, and Wang et al. (2022) assumes a density estimation oracle that controls the error between the estimated model and the real one. However, both algorithms use optimistic planning, which is computationally inefficient. In comparison, our algorithm achieves optimism by a UCB-type algorithm, and the only necessary oracle is MLE, which is more amenable to computation. To ensure short memory, Wang et al. (2022) make a constant past sufficiency assumption, which means that the trajectory density from the past L steps could determine the current state distribution. Therefore the sample complexity has an exponential dependence on that constant. As a comparison, our result for γ-observable POMDPs is more general and achieves better sample complexity. In Table 1, we compare our work to all prior works that achieve sample efficient learning in γ-observation POMDPs.

3. Preliminaries

Notations For any natural number $n \in \mathbb{N}$, we use $[n]$ to denote the set $\{1, \cdots, n\}$. For vectors we use $\|\cdot\|_p$ to denote ℓ_p-norm, and we use $\|x\|_A$ to denote $\sqrt{x^\top Ax}$. For a set $\mathcal{S}$ we use $\Delta(\mathcal{S})$ to denote the set of all probability distributions on $\mathcal{S}$. For an operator $\mathbb{O} : \mathcal{S} \to \mathbb{R}$ and $b \in \Delta(\mathcal{S})$, we use $\mathbb{O}b : \mathcal{O} \to \mathbb{R}$ to denote $\int_{\mathcal{S}} \mathbb{O}(o \mid s)b(s)\,\mathrm{d}s$. For two series $\{a_n\}_{n\geq1}$ and $\{b_n\}_{n\geq1}$, we use $a_n \leq O(b_n)$ to denote that there exists $C > 0$ such that $a_n \leq C \cdot b_n$.

POMDP Setting In this paper we consider a finite horizon *partially observable Markov decision process (POMDP)* $\mathcal{P}$, which can be specified as a tuple $\mathcal{P} = \big(\mathcal{S}, \mathcal{A}, H, \mathcal{O}, d_0, \{r_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H, \{\mathbb{O}_h\}_{h=1}^H\big)$. Here $\mathcal{S}$ is the state space, $\mathcal{A}$ is a finite set of actions, $H \in \mathbb{N}$ is the episode length, $\mathcal{O}$ is the set of observations, d_0 is the known initial distribution over states. For a given $h \in [H]$, $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the transition kernel and $r_h : \mathcal{O} \to [0, 1]$ is the reward function at step h. For each $a \in \mathcal{A}$, we abuse the notation and use $\mathbb{P}_h(a)$ to denote the probability transition kernel over the state space conditioned on action a in step h. $\mathbb{O}_h : \mathcal{S} \to \Delta(\mathcal{O})$ is the observation distribution at step h, where for $s \in \mathcal{S}, o \in \mathcal{O}$, $\mathbb{O}_h(o \mid s)$ is the probability of observing o while in state s at step h. We denote $r = (r_1, \ldots, r_H)$, $\mathbb{P} = (\mathbb{P}_1, \cdots, \mathbb{P}_H)$ and $\mathbb{O} = (\mathbb{O}_1, \cdots, \mathbb{O}_H)$.

We remark that *Markov Decision Process* (MDP) is a special case of POMDP, where $\mathcal{O} = \mathcal{S}$ and $\mathbb{O}_h(o \mid s) = \mathbb{1}\{o = s\}$ for all $h \in [H]$, $o \in \mathcal{O}$ and $s \in \mathcal{S}$.

Low-rank transition kernel In our work, we focus on the problem where the underlying transition of the environment has a low-rank structure.

Definition 1 (Low-rank transition (Agarwal et al., 2020; Uehara et al., 2021b)). *A transition kernel $\mathbb{P}_h : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ admits a low-rank decomposition of dimension d if there* *exists two mappings $\omega_h^*: \mathcal{S} \to \mathbb{R}^d$, and $\psi_h^*: \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ such that $\mathbb{P}_h(s' \mid s,a) = \omega_h^*(s')^\top \psi_h^*(s,a)$.*

Extended POMDP For notation convenience in the proof, we define the extended POMDP to allow $h < 0$. Specifically, we will extend a POMDP from step $h = 3 - 2L$ for a suitable choice of L. This particular choice is for the proof only and does not affect the execution of the algorithm in any way. The agent's interaction with the extended POMDP starts with an initial distribution d_0. For $s, s' \in \mathcal{S}$, $a \in \mathcal{A}$ and $h \leq 0$, we define $\phi_h^*(s,a) = d_0$, $\mu(s') = d_0(s')e_1$. Hence for a fixed constant $L > 0$, all the dummy action and observation sequences $\{o_{3-2L}, a_{3-2L}, \cdots, o_0, a_0\}$ leads to the same initial state distribution d_0. A general policy π is a tuple $\pi = (\pi_{3-2L}, \ldots, \pi_H)$, where $\pi_h : \mathcal{A}^{2L+h-3} \times \mathcal{O}^{2L+h-2} \to \Delta(\mathcal{A})$ is a mapping from histories up to step h, namely tuples $(o_{3-2L:h}, a_{3-2L:h-1})$, to actions. For L we will denote the collection of histories up to step h by $\mathcal{H}_h := \mathcal{A}^{2L+h-3} \times \mathcal{O}^{2L+h-3}$ and the set of policies by Π^{gen}, meaning that $\Pi^{\mathrm{gen}} = \Delta(\prod_{h=1}^H \mathcal{A}^{\mathcal{H}_h})$. For $z_h = (o_{h-L+1:h}, a_{h-L+1:h-1})$, we denote $z_{h+1} = c(z_h, a_h, o_{h+1}) = (o_{h-L+2:h+1}, a_{h-L+2:h})$. For any policy π, $h \in [H]$, and positive integer n, we use $\pi_h \circ_n U(\mathcal{A})$ to denote the policy that $(\pi_h \circ_n U(\mathcal{A}))_i = \pi_i$ for $i \leq h - n$ and $(\pi_h \circ_n U(\mathcal{A}))_i = U(\mathcal{A})$ for $i \geq h - n + 1$, which takes the first $h - n$ actions from π and takes remaining actions uniformly.

L-memory policy class In our work, we consider L-memory policies. For all $h \in [H]$, let $\mathcal{Z}_h = \mathcal{O}^L \times \mathcal{A}^{L-1}$. An element $z_h \in \mathcal{Z}_h$ is represented as $z_h = (o_{h+1-L:h}, a_{h+1-L:h-1})$. A L-memory policy is a tuple $\pi = (\pi_{3-2L}, \ldots, \pi_H)$, where $\pi_h : \mathcal{Z}_h \to \Delta(\mathcal{A})$. We define the set of L-memory policy by Π^L. We remark that $\Pi^L \subset \Pi^{\mathrm{gen}}$.

We define the value function for π at step 1 by $V_1^{\pi,\mathcal{P}}(o_1) = \mathbb{E}_{o_{2:H}\sim\pi}^{\mathcal{P}}[\sum_{h=1}^H r_h(o_h) \mid o_1]$, namely as the expected reward received by following π.

4. Warmup: Low-rank POMDPs with $L-$step decodability

To begin with, let us use the L-step decodable POMDP setting to motivate our algorithm.

Assumption 2 (Low-Rank L-step decodability)**.** *For all $h \in [H]$, $s_{h+1} \in \mathcal{S}$ and $(z_h, a_h) \in \mathcal{Z}_h \times \mathcal{A}$. There exists ϕ_h^* such that $\mathbb{P}_h(s_{h+1} \mid z_h, a_h) = \phi_h^*(z_h, a_h)^\top \omega_h^*(s_{h+1})$, where ω_h^* is the same function as given in Definition 1.*

Note that Assumption 2 is more general than the decodability assumption made in recent works (Efroni et al., 2022a; Uehara et al., 2022a). They assume that a suffix of length L of the history suffices to predict the latent state, i.e., there is some decoder $\mathbf{x} : \mathcal{Z}_i \to \mathcal{S}, i \in [H]$ such that $s_h = \mathbf{x}(z_h)$. It leads to our assumption that $\mathbb{P}(s_{h+1} \mid z_h, a_h) = \psi_h^*(\mathbf{x}(z_h), a_h)^\top \omega_h^*(s_{h+1})$, but not vice versa.

With Assumption 2, for any $z_h = (o_{h+1-L:h}, a_{h+1-L:h-1})$, a_h and o_{h+1}, we have

$$\mathbb{P}_h(o_{h+1} \mid o_{h+1-L:h}, a_{h+1-L:h}) \qquad (1)$$
$$= \Big[\int_{\mathcal{S}'} \omega_h^*(s')^\top \mathbb{O}_{h+1}(o_{h+1}^* \mid s')\, \mathrm{d}\, s'\Big] \phi_h^*(z_h, a_h).$$

For any $o_h \in \mathcal{O}$, we denote

$$\mu_h^*(o_h) = \int_{\mathcal{S}'} \omega_h^*(s')^\top \mathbb{O}_h^*(o_h \mid s')\, \mathrm{d}\, s' \qquad (2)$$

When considering state spaces of arbitrary size, it is necessary to apply function approximation to generalize across states. Representation learning is a natural way to allow such generalization by granting the agent access to a function class of candidate embeddings. Towards this end, we make a commonly used realizability assumption that the given function class contains the true function.

Assumption 3 (Realizability)**.** *For all $h \in [H]$, there exists a known model class $\mathcal{F} = \{(\mathbb{O}_h, \omega_h, \phi_h) : \mathbb{O}_h \in \mathcal{G}, \omega_h \in \Omega, \phi_h \in \Phi\}_{h=1}^H$, where $\mathbb{O}_h^* \in \mathcal{G}$, $\omega_h^* \in \Omega$ and $\phi_h^* \in \Phi$ for all $h \in [H]$.*

We compare this assumption to the work of Efroni et al. (2022a), which also studies in L-decodable POMDPs. Their learning in the function approximation sets assumes that the agent could get access to a function class that contains the Q^* function. Our assumption is more realizable since it is easier to have access to the transition class than the class of Q^* function.

4.1. Algorithm Description

In this section, we present our algorithm Partially Observable Representation Learning-based Reinforcement Learning (PORL^2).

In Algorithm 1, the agent operates in the underlying POMDP environment $\mathcal{P}$. In each episode, the agent takes three steps to update the new policy: (i) using a combination of the latest policy π and uniform policy $U(\mathcal{A})$ to collect data for each time step $h \in [H]$, (ii) calling the MLE oracle to learn and updating the representation $\widehat{\phi}_h$, and (iii) calculating an exploration bonus $\widehat{b}_h$ and applying Least Square Value Iteration for L-step Low-Rank POMDPs (LSVI-LLR, cf. Algorithm 2) to update our policy with the combined reward $r + \widehat{b}$.

The data collection process has two rounds. In the first round, the agent rollouts the current policy π^{k-1} for the first

Algorithm 1 Partially Observable Representation Learning for L-decodable POMDPs (PORL^2-decodable)

Require: Representation classes $\{\mathcal{F}_h\}_{h=0}^{H-1}$, parameters K, α_k, λ_k
1: Initialize policy $\pi^0 = \{\pi_0, \dots, \pi_{H-1}\}$ to be arbitrary policies and replay buffers $\mathcal{D}_h = \varnothing$, $\mathcal{D}'_h = \varnothing$ for all h.
2: **for** $k \in [K]$ **do**
3: Data collection from π^{k-1}, $\forall h \in [H]$, $\tau_h \sim d_{\mathcal{P}}^{\pi_h^{k-1} \circ_L U(\mathcal{A})}$; $\mathcal{D}_h = \mathcal{D}_h \cup \{\tau_h\}$, $\tilde{\tau}_h \sim d_{\mathcal{P}}^{\pi_h^{k-1} \circ_{2L} U(\mathcal{A})}$; $\mathcal{D}'_h = \mathcal{D}'_h \cup \{\tilde{\tau}_h\}$.
4: Learn representations for all $h \in [H]$: $(\widehat{\mathbb{O}}_h^k, \widehat{\omega}_h^k, \widehat{\phi}_h^k) = \mathrm{argmax}_{(\mathbb{O}_h, \omega_h, \psi_h) \in \mathcal{F}} \mathbb{E}_{\mathcal{D} \cup \mathcal{D}'}[\log \phi_h(z_h, a_h)^\top \mu_h(o_{h+1})]$, where μ_h is computed by (2).
5: Define exploration bonus for all $h \in [H]$: $\widehat{b}_h^k(z,a) = \min\left\{\alpha_k \sqrt{\widehat{\phi}_h^k(z,a)^\top \Sigma_h^{-1} \widehat{\phi}_h^k(z,a)}, 2\right\}$, with $\Sigma_h := \sum_{z \sim \mathcal{D}_h} \widehat{\phi}_h^k(z,a) \widehat{\phi}_h^k(z,a)^\top + \lambda_k I$.
6: Set π^k as the policy returned by calling Algorithm 2: LSVI-LLR($\{r_h + \widehat{b}_h^k\}_{h=0}^{H-1}, \{\widehat{\phi}_h^k\}_{h=0}^{H-1}, \{\widehat{\mu}_h^k\}_{h=0}^{H-1}, \{\mathcal{D}_h \cup \mathcal{D}'_h\}_{h=0}^{H-1}, \lambda_k$).
7: **end for**
8: **Return** $\pi^0, \cdots, \pi^K$

$h - L$ and takes $U(\mathcal{A})$ for the remaining steps to collect data. In the second round, the agent rollouts π^{k-1} for the first $h - 2L$ and takes $U(\mathcal{A})$ for the remaining steps. After collecting new data and concatenating it with the existing data, the agent learns the representation by calling the MLE oracle on the historical dataset (Line 3). Then, we set the exploration bonus based on the learned representation (Line 4), and we update the policy with the learned representations and bonus-enhanced reward (Line 5).

To update our policy, we apply LSVI-LLR – an adaptation of the classic LSVI algorithm to L-step low-rank POMDPs. For a given reward r and a model (μ, ϕ), the probability $\mathbb{P}_h^{\mathcal{P}}(c(z,a,o') \mid z,a) = \mu_h(o')^\top \phi_h(z,a)$. Therefore, we have $Q_h(z,a) = r_h(o) + \sum_{o' \in \mathcal{D}} \phi_h(z,a)^\top \mu_h(o'_{h+1}) V_{h+1}(z_{h+1})$, where $z_{h+1} = c(z_h, a_h, o'_h)$. After inductively computing the Q-function, we can out the greedy policy $\pi_h^t = \mathrm{argmax}_a Q_h(z_h, a)$.

Remark 4 (Computation)**.** *Regarding the computational cost, our algorithm only requires calling MLE computation oracle H times in every iteration. Optimism is achieved by adding a bonus to the reward function, which takes $O(Hd^2)$ flops in each iteration to compute with the Sherman-Morrison formula. Importantly, we avoid the optimistic planning procedure that requires iterating over the whole function class $\mathcal{F}$. Then the time complexity is dominated by the LSVI-LLR step 5 $\pi_h(z) = \mathrm{argmax}_{a \in \mathcal{A}} Q_h(z,a)$ for*

Algorithm 2 Least Square Value Iteration for L-step Low-Rank POMDPs (LSVI-LLR)

Require: $\{r_h\}_{h=0}^{H-1}$, features $\{\phi_h\}_{h=0}^{H-1}, \{\mu_h\}_{h=0}^{H-1}$, datasets $\{\mathcal{D}_h\}_{h=0}^{H-1}$, regularization λ.
1: Initialize $V_H(z) = 0$ for any $z \in \mathcal{Z}$.
2: **for** $h = H-1 \to 1$ **do**
3: For $(z_h, a_h) \in \mathcal{Z}_h \times \mathcal{A}_h$, set

$$Q_h(z_h, a_h) = r_h(o_h) + \sum_{o_{h+1} \in \mathcal{D}} \phi_h(z_h, a_h)^\top \mu_h(o_{+1}) \cdot V_{h+1}(c(z_h, a_h, o_{h+1})),$$

where $z_h = (o_{h-L+1:h}, a_{h-L+1:h-1})$.
4: Set $V_h(z) = \max_{a \in \mathcal{A}} Q_h(z,a)$.
5: Set $\pi_h(z) = \mathrm{argmax}_{a \in \mathcal{A}} Q_h(z,a)$.
6: **end for**
7: **Return** $\pi = \{\pi_0, \cdots, \pi_{H-1}\}$.

all $(z,a) \in \mathcal{D}^k$, which causes a $O(AHd^2K)$ running time in every iteration and a total $O(AHd^2K^2)$ running time. Therefore, our algorithm is much more amendable to a practical implementation.

4.2. Analysis

We have the following guarantee of our Algorithm 1.

Theorem 5 (Sample complexity of PORL^2-L-decodable)**.** *Under Assumption 2 and Assumption 3, for fixed $\delta, \epsilon \in (0,1)$, and let $\overline{\pi}$ be a uniform mixture of $\pi^0, \dots, \pi^{K-1}$. By setting the parameters as*

$$\alpha_k = \tilde{\Theta}(\sqrt{k|\mathcal{A}|^L \zeta_k + \lambda_k d + k\zeta_k}),$$
$$\lambda_k = \Theta(d \log(|\mathcal{F}|k/\delta)),$$

with prob. at least $1-\delta$, we have $V_1^{\mathcal{P},\pi^*,r} - V_1^{\mathcal{P},\overline{\pi},r} \le \epsilon$, after

$$H \cdot K = \mathcal{O}\left(\frac{H^5 |\mathcal{A}|^{2L} d^4 \log(d|\mathcal{F}|/\delta)}{\epsilon^2}\right).$$

samples, where π^* is the optimal policy of Π^{gen}.

Theorem 5 indicates that the sample complexity of PORL^2 only depends polynomially on the rank d, the history step L, horizon H, the size of the effective action space for L-memory policy $|\mathcal{A}|^L$, and the statistical complexity of the function class $\log(|\mathcal{F}|)$. In particular, Theorem 5 avoids direct dependency on the size of the state or observation space. Specifically, we emphasized the term $|\mathcal{A}|^{2L}$, which comes from doing importance sampling for $2L$ times in line 3 of Algorithm 1. Our sample complexity also matches the regret bound of Liu et al. (2022a), Theorem 7.

5. Low-rank POMDPs with γ - Observability

In this section, we move on to the observability POMDP setting

Assumption 6 (Golowich et al. (2022a;b); Even-Dar et al. (2007))**.** *Let $\gamma > 0$. For $h \in [H]$, let $\mathbb{O}_h$ be the operator with $\mathbb{O}_h(\cdot \mid s)$, indexed by states s. We say that the operator $\mathbb{O}_h$ satisfies γ-observability if for each h, for any distributions b, b' over states, $\|\mathbb{O}_h b - \mathbb{O}_h b'\|_1 \geq \gamma \|b - b'\|_1$. A POMDP satisfies γ-observability if all $h \in [H]$ of satisfy γ - observability.*

Assumption 6 implies that the operator $\mathbb{O}_h : \Delta(S) \to \Delta(O)$ is an injection. We use τ_i to denote $(o_{3-2L:i},\ a_{3-2L:i-1})$. In addition, we make the same realizability assumption (Assumption 3) as in the decodable setting.

For γ-observable low-rank POMDPs, we present the assumption that are commonly adopted in the literature to avoid challenges associated with reinforcement learning with function approximation (Efroni et al., 2022b). We state the function approximation and computational oracles below.

Assumption 7. *There exists a known model class $\mathcal{F} = \{(\mathbb{O}_h, \omega_h, \psi_h) : \mathbb{O}_h \in \mathcal{G}, \omega \in \Omega, \psi_h \in \Psi_h\}_{h=1}^H$, where $\mathbb{O}_h^* \in \mathcal{G}$, $\omega_h^* \in \Omega$ and $\psi_h^* \in \Psi$ for all $h \in [H]$. Recall that $\mathbb{P}_h(o_{h+1} \mid s_h, a_h) = \mathbb{O}_{h+1}^*(o_{h+1} \mid s_{h+1})\omega_{h+1}^*(s_{h+1})^\top \psi_h^*(s_h, a_h)$.*

Compared to Assumption 3, we assume the model class contains the transition information of the latent state in this assumption.

5.1. The approximated MDP $\mathcal{M}$ with L-structure

It has been shown that when γ-observability holds, the POMDP can be approximated by an MDP whose state space is $\mathcal{Z} = \mathcal{O}^L \times \mathcal{A}^{L-1}$ (Uehara et al., 2022b). In particular, for a probability transition kernel $\mathbb{P}_h(z_{h+1} \mid z_h, a_h)$ (for $z_h, z_{h+1} \in \mathcal{Z}$) and a reward function $r = (r_1, \ldots, r_H), r_h : \mathcal{Z} \to \mathbb{R}$, we will consider MDPs of the form $\mathcal{M} = (\mathcal{Z}, \mathcal{A}, H, r, \mathbb{P})$. For such an $\mathcal{M}$, we say that $\mathcal{M}$ has L-structure if: the transitions $\mathbb{P}_h(\cdot \mid z_h, a_h)$ have the following property, for $z_h \in \mathcal{Z}, a_h \in \mathcal{A}$: writing $z_h = (o_{h-L+1:h}, a_{h-L:h-1}), \mathbb{P}_h(z_{h+1} \mid z_h, a_h)$ is nonzero only for those z_{h+1} of the form $z_{h+1} = (o_{h-L+2:h+1}, a_{h-L+1:h})$, where $o_{h+1} \in \mathcal{O}$.

For a low-rank POMDP $\mathcal{P}$, o_{h+1} can only be predicted by the whole memory $\{o_1,\ a_1, \cdots,\ o_h, a_h\}$. The main observation in (Uehara et al., 2022b) is that o_{h+1} can instead be approximated predicted by the L-memory $\{o_{h+1-L},\ a_{h+1-L}, \cdots,\ o_h, a_h\}$ with an error bound ϵ_1, given the memory length is at least $L = O(\gamma^{-4} \log(d/\epsilon_1))$. In other words, there exists an approximated **low-rank MDP $\mathcal{M}$ with L-structure** that is close to $\mathcal{P}$. For any $\mathcal{P} = (\mathbb{O}, \omega, \psi)$, we can construct an approximated MDP $\mathcal{M} = \{(\mu_h, \phi_h)\}_{h=1}^H$, where $(\phi, \mu) = q(\mathbb{O}, \omega, \psi)$ for an explicit function q. The analytical form of q is not important for our discussion and is deferred to Appendix D.1. This approximated MDP $\mathcal{M}$ satisfies that

$$\mathbb{P}_h^{\mathcal{P}}(o_{h+1} \mid z_h, a_h) = \mu_{h+1}^\top(o_{h+1})\phi_h(z_h, a_h). \tag{3}$$

At the same time, the POMDP $\mathcal{P}$ satisfies that

$$\mathbb{P}_h^{\mathcal{P}}(o_{h+1} \mid \tau_h, a_h) = \mu_{h+1}^\top(o_{h+1})\xi_h(\tau_h, a_h), \tag{4}$$

where the definition of ξ is also defferred to Appendix D.1.

The constructed $\mathcal{M}$ retains the structure of low-rank POMDP, and we have the following proposition:

Proposition 8. *For any $\epsilon_1 > 0$, there exists an L-structured MDP $\mathcal{M}$ with $L = O(\gamma^{-4} \log(d/\epsilon_1))$, such that for all $\pi \in \Pi^{gen}$ and $h \in [H]$,*

$$\mathbb{E}_{a_{1:h}, o_{2:h} \sim \pi} \|\mathbb{P}_h^{\mathcal{M}}(o_{h+1} \mid z_h, a_h) - \mathbb{P}_h^{\mathcal{P}}(o_{h+1} \mid o_{1:h}, a_{1:h})\|_1 \leq \epsilon_1.$$

By Proposition 8, we have that the conditional probability $\mathbb{P}^{\mathcal{P}}(o \mid z, a)$ is approximately low rank. Now, we can define the value function under $\mathcal{M}$ as $V^{\pi, \mathcal{M}, r}(o_1) = \mathbb{E}_h^{\pi, \mathcal{M}}[\sum_{i=h}^H r_i \mid o_1]$. With Proposition 8, we can prove that for a L-memory policy π, the value function of π in $\mathcal{M}$ can effectively approximate the value function under $\mathcal{P}$.

Lemma 9. *With $\mathcal{M}$ defined in (3), for any policy $\pi \in \Pi^{\mathrm{gen}}$, we have*

$$|V_1^{\pi, \mathcal{P}, r}(o_1) - V_1^{\pi, \mathcal{M}, r}(o_1)| \leq \frac{H^2 \epsilon_1}{2}.$$

A direct implication of Lemma 9 is that a near-optimal policy of $\mathcal{M}$ is also near-optimal in $\mathcal{P}$. Hence we only need to focus on finding the optimal policy in Π^L, the set of Markovian policy in $\mathcal{M}$.

5.2. Algorithm and Analysis

PORL^2 for γ-observable POMDPs is presented in Algorithm 3. It is almost identical to Algorithm 1 except for the representation learning step, where in the MLE we maximize $\log \mathbb{P}_h^{\mathcal{P}}(o_{h+1} \mid \tau_h, a_h)$. Note that we have $\mathbb{P}_h^{\mathcal{P}}(o_{h+1} \mid \tau_h, a_h) = \xi_h(\tau_h, a_h)^\top \mu_{h+1}(o_{h+1})$ as shown in (4).

Regarding computational cost, similar to Algorithm 1, Algorithm 3 only requires a total of HK calls to the MLE oracle and additional polynomial time complexity for executing the LSVI-LLR planner. The main difference lies in the additional transformation q that needs to be performed when calculating the log-likelihood $\log \mathbb{P}_h^{\mathcal{P}}(o_{h+1} \mid \tau_h, a_h)$. For instance, when using gradient-based optimization to solve

Algorithm 3 Partially Observable Representation Learning for γ-observable POMDPs (PORL^2-γ-observable)

Require: Representation classes $\{\mathcal{F}_h\}_{h=0}^{H-1}$, parameters K, α_k, λ_k
1: Initialize policy $\pi^0 = \{\pi_0, \ldots, \pi_{H-1}\}$ to be arbitrary policies and replay buffers $\mathcal{D}_h = \varnothing$, $\mathcal{D}'_h = \varnothing$ for all h.
2: **for** $k \in [K]$ **do**
3: Data collection from π^{k-1}, $\forall h \in [H]$, $\tau_h \sim d_{\mathcal{P}}^{\pi_h^{k-1} \circ_L U(\mathcal{A})}$; $\mathcal{D}_h = \mathcal{D}_h \cup \{\tau_h\}$, $\tilde{\tau}_h \sim d_{\mathcal{P}}^{\pi_h^{k-1} \circ_{2L} U(\mathcal{A})}$; $\mathcal{D}'_h = \mathcal{D}'_h \cup \{\tilde{\tau}_h\}$.
4: Learn representations for all $h \in [H]$: $(\mathbb{O}^k, \omega^k, \psi^k) = \operatorname{argmax}_{(\mathbb{O}_h, \omega_h, \psi_h) \in \mathcal{F}} \mathbb{E}_{\mathcal{D} \cup \mathcal{D}'}[\log \xi_h(\tau_h, a_h)^\top \mu_{h+1}(o_{h+1})]$ by (4).
5: Learn the L-step feature: $(\widehat{\phi}^k, \widehat{\mu}^k) = q(\mathbb{O}^k, \omega^k, \psi^k)$
6: Define exploration bonus for all $h \in [H]$: $\widehat{b}_h^k(z, a) = \min\left\{\alpha_k \sqrt{\widehat{\phi}_h^k(z, a)^\top \Sigma_h^{-1} \widehat{\phi}_h^k(z, a)}, 2\right\}$, with $\Sigma_h := \sum_{z \sim \mathcal{D}_h} \widehat{\phi}_h^k(z, a) \widehat{\phi}_h^k(z, a)^\top + \lambda_k I$.
7: Set π^k as the policy returned by: LSVI-LLR($\{r_h + \widehat{b}_h^k\}_{h=0}^{H-1}, \{\widehat{\phi}_h^k\}_{h=0}^{H-1}, \{\widehat{\mu}_h^k\}_{h=0}^{H-1}, \{\mathcal{D}_h \cup \mathcal{D}'_h\}_{h=0}^{H-1}, \lambda_k$).
8: **end for**
9: **Return** $\pi^0, \cdots, \pi^K$

the MLE problem, the total number of such calculations performed is usually small and does not scale directly with the size of the function class.

The next theorem shows that Algorithm 3 achieves the same sample complexity in γ-observable POMDPs as in Theorem 5.

Theorem 10 (Pac Bound of PORL^2-γ- observability)**.** *Under Assumption 6 and Assumption 3, Let $\delta, \epsilon \in (0, 1)$ be given, and let $\overline{\pi}$ be a uniform mixture of $\pi^0, \ldots, \pi^{K-1}$. By setting the parameters as*

$$\alpha_k = \tilde{\Theta}(\sqrt{k|\mathcal{A}|^L \zeta_k + \lambda_k d}),\ \lambda_k = \Theta(d \log(|\mathcal{F}|k/\delta)),$$
$$\epsilon_1 = \Theta(\epsilon / (H^2 d^{1/2} \gamma^{-4} \log(1/\epsilon) \log(dHA|\mathcal{F}|/\delta)^{1/2})),$$
$$L = \Theta(\gamma^{-4} \log(d/\epsilon_1)),\ \zeta_k = \Theta(\log(|\mathcal{F}|k/\delta)/k),$$

with probability at least $1 - \delta$, we have $V^{\mathcal{P}, \pi^*, r} - V^{\mathcal{P}, \overline{\pi}, r} \leq \epsilon$, after

$$H \cdot K = \mathcal{O}\left(\frac{H^5 |\mathcal{A}|^{2L} d^4 \log(d|\mathcal{F}|/\delta)}{\epsilon^2}\right).$$

episodes of interaction with the environment, where π^* is the optimal policy of Π^{gen}.

We remark on the $|\mathcal{A}|^{2L}$ term, which comes from the importance sampling in Algorithm 3, matches the regret bound in Liu et al. (2022a), and the sample complexity in (Golowich et al., 2022a) for γ-observable POMDP. In addition, it has been shown in (Golowich et al., 2022b), Theorem 6.4 that this sample complexity is necessary for any computational-efficient algorithm.

6. Highlight of the Analysis

In this section, we highlight the critical observations in our analysis of Theorem 5 and Theorem 10.

MLE guarantee. The following lemma upper-bounds the reconstruction error with the learned features at any iteration of PORL^2.

Lemma 11 (MLE guarantee)**.** *Set $\lambda_k = \Theta(d \log(|\mathcal{F}|k/\delta))$, for any time step $h \in [H]$, denote ρ_h as the joint distribution for (w, a) in the dataset $\mathcal{D}$ of step k, with probability at least $1 - \delta$ we have*

$$\mathbb{E}_{w,a \sim \rho}[\|\mathbb{P}_h^{\widehat{\mathcal{P}}}(\cdot \mid w, a)^\top - \mathbb{P}_h^{\mathcal{P}}(\cdot \mid w, a)^\top\|_1^2] \leq \zeta_k, \quad (5)$$

recall that $\zeta_k = O(\log(|\mathcal{F}|k/\delta)/k)$. Here w is the trajectory τ in the decodable case (Algorithm 1), and w is the state z of the approximated MDP $\mathcal{M}$ in the γ-observable case (Algorithm 3).

L-step Back Inequality. It can be observed that the MLE guarantee ensures the expectation of model estimation error scales as $\tilde{O}(1/k)$ scale under the occupancy distribution of the average policy of $\pi^0, ..., \pi^{k-1}$. However, to estimate the performance of the policy π^k, we must perform a distribution transfer from the distribution induced by $\pi^0, ..., \pi^{k-1}$ to the distribution induced by π^k. To deal with this issue, we generalize the one-step-back technique in Uehara et al. (2021b); Agarwal et al. (2020) and propose a novel L-step-back inequality to handle L-memory policies. This L-step back inequality,i.e., moving from h to $h - L$, leverages the bilinear structure in $\mathbb{P}_{h-L}^{\mathcal{P}}$.

Lemma 12 (L-step back inequality)**.** *For any $h \in [H]$ and $k \in [K]$, consider function g_h that satisfies $g_h \in \mathcal{Z}_h \times \mathcal{A}_h \to \mathbb{R}$, s.t. $\|g_h\|_\infty \leq B$. Then, for any policy π, we have*

$$\begin{aligned}
&\mathbb{E}_\pi^{\mathcal{P}}[g(z_h, a_h)] \\
&\leq \mathbb{E}_{(z,a)_{h-L-1} \sim \pi}^{\mathcal{P}}\left[\|\phi(z_{h-L-1}, a_{h-L-1})\|_{\Sigma_{\beta_{h-L-1}^k, \phi_{h-L-1}}}^{-1}\right] \\
&\qquad \cdot \sqrt{|A|^L k \cdot \mathbb{E}_{(\tilde{z}_h, \tilde{a}_h) \sim \gamma_h}\{[g(\tilde{z}_h, \tilde{a}_h)]^2\} + B^2 \lambda_k d},
\end{aligned}$$

where we denote $\gamma_h^k = 1/k \sum_{i=0}^{k-1} d_h^{\pi^i}(z, a)$ and $\beta_h^k = 1/k \sum_{i=0}^{k-1} d_h^{\pi^i \circ_{2L} U(A)}(z, a)$ as the mixture state-action distribution, and $\Sigma_{\beta_{h-L-1}^k, \phi} = k \mathbb{E}_{z,a \sim \beta_{h-L-1}} \phi_{h-L-1}(z, a) \phi_{h-L-1}(z, a)^\top + \lambda_k I$ as the regularized covariance matrix under the representation ϕ.

Almost Optimism. For $\pi^k = \text{LSVI-LLR}(r+\widehat{b}, \widehat{\phi}, \{\mathcal{D}^k \cup \mathcal{D}^{k\prime}, \lambda_k\})$, we can prove that the value function of π^k is almost optimism at the initial state distribution.

Lemma 13 (Almost Optimism at the Initial State Distribution)**.** *Using the parameters of Theorem 10, with probability $1-\delta$, we have for all iterations $k \in [K]$,*

$$V^{\pi^*, \widehat{\mathcal{P}}_k, r+\widehat{b}^k} - V^{\pi^*, \mathcal{P}, r} \geq -\sigma_k,$$

where $\sigma_k = \mathcal{O}(Ld(A^L \log(dk|\mathcal{F}|/\delta)/k)^{1/2})$.

Then, using Lemma 13 and the simulation lemma, we can establish an upper bound on the policy regret for π^k, we have:

$$\begin{aligned} & V_1^{\pi^*, \mathcal{P}, r} - V_1^{\pi^k, \mathcal{P}, r} \\ & \quad \leq V^{\pi^k, \widehat{\mathcal{P}}, r+\widehat{b}^k} - V^{\pi^k, \mathcal{P}, r} + \sigma_k \\ & \quad \leq \sum_{h=0}^{H-1} \mathbb{E}_{z,a}\Big[\widehat{b}_h^k + \mathbb{E}_{o' \sim \mathbb{P}_h^{\mathcal{P}}(z,a)} V_{h+1}^{\pi^k, \mathcal{P}, \widehat{b}_h^k}(z_{h+1}) \\ & \qquad\quad - \mathbb{E}_{o' \sim \mathbb{P}_h^{\widehat{\mathcal{P}}}(z,a)} V_{h+1}^{\pi^k, \mathcal{P}, \widehat{b}_h^k}(z_{h+1})\Big] + \sigma_k, \end{aligned} \tag{6}$$

where $z_{h+1} = c(z_h, a_h, o')$.

Finally, we give an upper bound for (6). We adopt the idea of the moment-matching policy in (Efroni et al., 2022a; Uehara et al., 2022a), which analyze the latent state of the past L-steps. For any g, $\mathbb{E}_\pi^{\mathcal{P}}[g(z_h, a_h)]$ can be written in a bilinear form

$$\begin{aligned} \Big\langle & \mathbb{E}^{\mathcal{P}}_{z_{h-L}, a_{h-L} \sim \pi} \phi^\top(z_{h-L}, a_{h-L}), \int_{s_{h-L+1} \in \mathcal{S}} \omega(s_{h-L+1}) \\ & \cdot \mathbb{E}^{\mathcal{P}}_{a_{h-L+1:h} \sim \mu^{\pi,h}}[g(z_h, a_h) \mid s_{h-L+1}] \,\mathrm{d}s_{h-L+1} \Big\rangle, \end{aligned}$$

where μ is the moment matching policy defined in (12).

Now we denote $\gamma_h^k = 1/k \sum_{i=0}^{k-1} d_h^{\pi^i}(z,a)$ to represent the mixture state-action distribution. Additionally, we define $\Sigma_{\gamma_h^k, \phi^*} = k\mathbb{E}_{z,a \sim \gamma_h^k} \phi^*(z,a)\phi^*(z,a)^\top + \lambda_k I$ as the regularized covariance matrix under the ground truth representation ϕ^*. To derive an upper bound for the aforementioned bilinear form, we can apply the Cauchy-Schwartz inequality within the norm induced by $\Sigma_{\gamma_h^k, \phi^*}$ as follows:

$$\begin{aligned} & \Big\langle \mathbb{E}^{\mathcal{P}}_{z_{h-L}, a_{h-L} \sim \pi} \phi_h^\top(z_{h-L}, a_{h-L}), \int_{s_{h-L+1} \in \mathcal{S}} \omega_h(s_{h-L+1}) \\ & \quad \cdot \mathbb{E}^{\mathcal{P}}_{a_{h-L+1:h} \sim \mu^{\pi,h}}[g(z_h, a_h) \mid s_{h-L+1}] \,\mathrm{d}s_{h-L+1} \Big\rangle \\ & \leq \mathbb{E}^{\widehat{\mathcal{P}}}_{z_{h-L}, a_{h-L} \sim \pi} \|\widehat{\phi}_h^\top(z_{h-L}, a_{h-L})\|_{\Sigma^{-1}_{\gamma_{h-L}, \widehat{\phi}_{h-L}}} \\ & \quad \cdot \Big\| \int_{s_{h-L+1} \in \mathcal{S}} \widehat{\omega}_h(s_{h-L+1}) \cdot \mathbb{E}^{\widehat{\mathcal{P}}}_{a_{h-L+1:h} \sim \mu^{\pi,h}}[g(z_h, a_h) \\ & \qquad \mid s_{h-L+1}] \,\mathrm{d}s_{h-L+1} \Big\|_{\Sigma_{\gamma_{h-L}, \widehat{\phi}_{h-L}}}, \end{aligned} \tag{7}$$

The first term in equation (7) is associated with the elliptical potential function. By employing the L-step back inequality, we can transform the second term in equation (7) into an expectation over the dataset distribution. This expectation can be controlled by leveraging the MLE guarantee.

Approximated Transition Error For the low-rank POMDPs with γ-observability, the main idea is to analyze the value function under the approximated MDP $\mathcal{M}$ instead of the real POMDP transition $\mathcal{P}$. We remark that is a novel technique in low-rank POMDPs with γ-observable assumption, which analyzes in a Markovian model with a small approximation error. The detailed proof of Theorem 10 can be found in Appendix D.

7. Experiments

We evaluate the performance of PORL^2 using the partially observed combination lock (pocomblock) as our benchmark, which is inspired by the combination lock benchmark introduced by Misra et al. (2019). Pocomblock consists of latent states accompanied by rich observations. Further details and specifics regarding the experiments can be found in Appendix E.

Next, we provide an overview of pocomblock. Pocomblock has three states: two good states and one bad state. When the agent is in a good state, it will remain in either of the two good states only if the correct action is taken; otherwise, it will transition to the bad state. Once the agent enters the bad state, it becomes impossible to exit. In the bad state, the agent receives a reward of zero.

Then we explain the emission kernel, when the time step h is odd, all the latent states generate rich observations and different states generate different observations. When h is even, one good state still has a rich observation space, while the other states' observations were absorbed by an absorbed observation. Hence at this time, it is unable to distinguish such two states by the current observation.

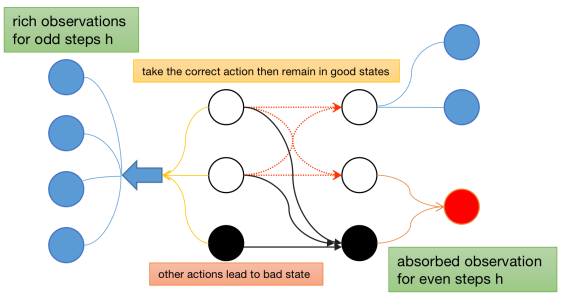

Figure 1. Visualization of pocomblock, where the blue area represents the rich observations obtained from the latent states. The black arrows illustrate the transition from the good states (depicted in white) to the bad state (depicted in black). Conversely, the red arrows indicate remaining in the good states by taking the correct action. Once the agent transitions to the bad state, it remains in that state for the entire episode, resulting in a failure to achieve the goal. Moreover, when the value of step h is even, there exists an absorbed observation that encompasses all the observations for both the bad state and one of the good states. Consequently, the agent is unable to distinguish between these two states solely based on a one-step observation during such time steps. Thus the name partially observed combination lock, (see Appendix E for details).

Comparision with BRIEE. We test PORL^2 and BRIEE (Zhang et al., 2022b) —the SOTA block MDP algorithm in pocomblock. We note that pocomblock is not a block MDP since the existence of the absorbed observation when h is even. The details and results can be found in Appendix E

Reproducibility. Our model and code can be found at https://github.com/icmlpomdpexpe/POMDPreplearn.

8. Conclusion

We presented a representation learning-based reinforcement learning algorithm for decodable and observable POMDPs, that achieves a polynomial sample complexity guarantee and is amendable to scalable implementation. Future works include empirically designing an efficient and scalable implementation of our algorithm and performing extensive empirical evaluations on public benchmarks, and theoretically extending our algorithm and framework to handle more general decision-making problems such as Predictive State Representations and beyond.

Acknowledgements. Mengdi Wang acknowledges the support by NSF grants DMS-1953686, IIS-2107304, CMMI-1653435, ONR grant 1006977, and C3.AI.

References

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.

Agarwal, A., Song, Y., Sun, W., Wang, K., Wang, M., and Zhang, X. Provable benefits of representational transfer in reinforcement learning, 2022. URL `https://arxiv.org/abs/2205.14571`.

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

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning*, pp. 463–474. PMLR, 2020a.

Ayoub, A., Jia, Z., Szepesvari, C., Wang, M., and Yang, L. F. Model-based reinforcement learning with value-targeted regression, 2020b. URL `https://arxiv.org/abs/2006.01107`.

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

Braziunas, D. Pomdp solution methods. *University of Toronto*, 2003.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Cassandra, A. R. A survey of pomdp applications. In *Working notes of AAAI 1998 fall symposium on planning with partially observable Markov decision processes*, volume 1724, 1998.

Chen, M., Frazzoli, E., Hsu, D., and Lee, W. S. Pomdp-lite for robust robot planning under uncertainty. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 5427–5433. IEEE, 2016.

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

Efroni, Y., Jin, C., Krishnamurthy, A., and Miryoosefi, S. Provable reinforcement learning with a short-term memory. *CoRR*, abs/2202.03983, 2022a. URL `https://arxiv.org/abs/2202.03983`.

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

Even-Dar, E., Kakade, S. M., and Mansour, Y. The value of observation for monitoring dynamic systems. In *IJCAI*, pp. 2474–2479, 2007.

Ghandali, R., Abooei, M., and Nezhad, F. A pomdp framework to find optimal inspection and maintenance policies via availability and profit maximization for manufacturing systems. *International Journal of Engineering*, 31 (12):2077–2084, 2018.

Golowich, N., Moitra, A., and Rohatgi, D. Learning in observable pomdps, without computationally intractable oracles. *arXiv preprint arXiv:2206.03446*, 2022a.

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

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

Guo, Z. D., Thakoor, S., Pslar, M., Pires, B. A., Altché, F., Tallec, C., Saade, A., Calandriello, D., Grill, J.-B., Tang, Y., et al. Byol-explore: Exploration by bootstrapped prediction. *arXiv preprint arXiv:2206.08332*, 2022.

Jafarnia-Jahromi, M., Jain, R., and Nayyar, A. Online learning for unknown partially observable mdps. *arXiv preprint arXiv:2102.12661*, 2021.

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

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

Katt, S., Oliehoek, F., and Amato, C. Bayesian reinforcement learning in factored pomdps. *arXiv preprint arXiv:1811.05612*, 2018.

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

Li, H., Liao, X., and Carin, L. Multi-task reinforcement learning in partially observable stochastic environments. *Journal of Machine Learning Research*, 10(5), 2009.

Liu, Q., Chung, A., Szepesvári, C., and Jin, C. When is partially observable reinforcement learning not scary? In *Conference on Learning Theory*, pp. 5175–5220. PMLR, 2022a.

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

Liu, Y., Liu, Q., Zhao, H., Pan, Z., and Liu, C. Adaptive quantitative trading: An imitative deep reinforcement learning approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 2128–2135, 2020.

Misra, D., Henaff, M., Krishnamurthy, A., and Langford, J. Kinematic state abstraction and provably efficient rich-observation reinforcement learning, 2019.

Misra, D., Henaff, M., Krishnamurthy, A., and Langford, J. Kinematic state abstraction and provably efficient rich-observation reinforcement learning. In *International conference on machine learning*, pp. 6961–6971. PMLR, 2020.

Modi, A., Chen, J., Krishnamurthy, A., Jiang, N., and Agarwal, A. Model-free representation learning and exploration in low-rank mdps. *CoRR*, abs/2102.07035, 2021. URL https://arxiv.org/abs/2102.07035.

Mossel, E. and Roch, S. Learning nonsingular phylogenies and hidden markov models. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pp. 366–375, 2005.

Mundhenk, M., Goldsmith, J., Lusena, C., and Allender, E. Complexity of finite-horizon markov decision process problems. *J. ACM*, 47(4):681–720, jul 2000. ISSN 0004-5411. doi: 10.1145/347476.347480. URL https://doi.org/10.1145/347476.347480.

Murphy, K. P. A survey of pomdp solution techniques. *environment*, 2(10), 2000.

Papadimitriou, C. H. and Tsitsiklis, J. N. The complexity of markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987a. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/3689975.

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

Roy, N., Gordon, G., and Thrun, S. Planning under uncertainty for reliable health care robotics. In *Field and Service Robotics*, pp. 417–426. Springer, 2006.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018. doi: 10.1126/science.aar6404. URL https://www.science.org/doi/abs/10.1126/science.aar6404.

Stooke, A., Lee, K., Abbeel, P., and Laskin, M. Decoupling representation learning from reinforcement learning. In Meila, M. and Zhang, T. (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9870–9879. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/stooke21a.html.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline rl in low-rank mdps, 2021a. URL https://arxiv.org/abs/2110.04652.

Uehara, M., Zhang, X., and Sun, W. Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*, 2021b.

Uehara, M., Sekhari, A., Lee, J. D., Kallus, N., and Sun, W. Provably efficient reinforcement learning in partially observable dynamical systems, 2022a. URL https://arxiv.org/abs/2206.12020.

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*, 2022b.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Embed to control partially observed systems: Representation learning with provable sample efficiency. *arXiv preprint arXiv:2205.13476*, 2022.

Yang, L. and Wang, M. Reinforcement learning in feature space: Matrix bandit, kernels, and regret bound. In *International Conference on Machine Learning*, pp. 10746–10756. PMLR, 2020.

Zanette, A., Cheng, C.-A., and Agarwal, A. Cautiously optimistic policy optimization and exploration with linear function approximation. In *Conference on Learning Theory*, pp. 4473–4525. PMLR, 2021.

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

Zhang, T., Ren, T., Yang, M., Gonzalez, J., Schuurmans, D., and Dai, B. Making linear mdps practical via contrastive representation learning. In *International Conference on Machine Learning*, pp. 26447–26466. PMLR, 2022a.

Zhang, X., Song, Y., Uehara, M., Wang, M., Agarwal, A., and Sun, W. Efficient reinforcement learning in block mdps: A model-free representation learning approach. *CoRR*, abs/2202.00063, 2022b. URL https://arxiv.org/abs/2202.00063.

A. Technical Lemma

In this section we introduce several lemmas which are useful in our proof.

Lemma 14 ((Uehara et al., 2021b)). *Consider the following process. For $n = 1, \cdots, N$, $M_n = M_{n-1} + G_n$ with $M_0 = \lambda_0 I$ and G_n being a positive semidefinite matrix with eigenvalues upper-bounded by 1 . We have that:*

$$2 \log \det(M_N) - 2 \log \det(\lambda_0 I) \geq \sum_{n=1}^{N} \mathrm{Tr}(G_n M_{n-1}^{-1}).$$

Lemma 15 (Elliptical Potential Lemma, (Uehara et al., 2021b)). *Suppose $\mathrm{Tr}(G_n) \leq B^2$, where G_n being a positive semidefinite matrix with eigenvalues upper-bounded by 1.*

$$2 \log \det(M_N) - 2 \log \det(\lambda_0 I) \leq d \log(1 + \frac{NB^2}{d\lambda_0}).$$

Next, we provide an important lemma to ensure the concentration of the bonus term in our algorithm. This lemma is proved in Lemma 39 of (Zanette et al., 2021).

Lemma 16 (Concentration of the Bonus). *Set $\lambda_n = \Theta(d \log(n|\mathcal{F}|/\delta))$ for any n. Let $\mathcal{D} = \{s_i, a_i\}_{i=0}^{n-1}$ be a stochastic sequence of data where $(s_i, a_i) \sim \rho_i$ where ρ_i can depend on the history of time steps $1, \ldots, i-1$. Let $\rho = \frac{1}{n}\sum_{i=0}^{n-1} \rho$ and define*

$$\Sigma_{\rho,\phi} = k\mathbb{E}_\rho[\phi(s,a)\phi^\top(s,a)] + \lambda_n I, \quad \widehat{\Sigma}_{n,\phi} = \sum_{i=0}^{n-1} \phi(s_i,a_i)\phi^\top(s_i,a_i) + \lambda_n I.$$

Then, with probability $1-\delta$, we have

$$\forall n \in \mathbb{N}^+, \forall \phi \in \Phi, c_1 \|\phi(s,a)\|_{\Sigma_{\rho,\phi}^{-1}} \leq \|\phi(s,a)\|_{\widehat{\Sigma}_{n,\phi}^{-1}} \leq c_2 \|\phi(s,a)\|_{\Sigma_{\rho,\phi}^{-1}}.$$

We also introduce the Simulation Lemma which is frequently used in RL literature. Its proof can be found in (Uehara et al., 2021b).

Lemma 17 (Simulation Lemma). *Given two MDPs $(P', r+b)$ and (P, r), for any policy π, we have:*

$$V^\pi_{P',r+b} - V^\pi_{P,r} = \sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim d^\pi_{P',h}}[b_h(s_h,a_h) + \mathbb{E}_{P'_h(s'_h|s_h,a_h)}[V^\pi_{P,r,h+1}(s'_h)] - \mathbb{E}_{P_h(s'_h|s_h,a_h)}[V^\pi_{P,r,h+1}(s'_h)]],$$

and

$$V^\pi_{P',r+b} - V^\pi_{P,r} = \sum_{h=1}^{H} \mathbb{E}_{(s_h,a_h)\sim d^\pi_{P,h}}[b_h(s_h,a_h) + \mathbb{E}_{P'_h(s'_h|s_h,a_h)}[V^\pi_{P,r+b,h+1}(s'_h)] - \mathbb{E}_{P_h(s'_h|s_h,a_h)}[V^\pi_{P,r+b,h+1}(s'_h)]].$$

We note that since both the occupancy measure and Bellman updates under $\widehat{P}$ are defined in the exact same way as if $\widehat{P}$ is a proper probability matrix, the classic simulation lemma also applies to $\widehat{P}$.

B. Proof of Theorem 5

First we define a few mixture notations that will be used extensively in the analysis. For any k, we define $\overline{\pi}^k$ to be $\sum_{i=0}^{k-1} \pi^k/k$. For a fixed (k,h), ρ_h^k is the distribution on $\mathcal{Z} \times \mathcal{A}$ induced by applying $\{\overline{\pi}_i^k\}_{i=1}^h$ and then do uniformly random action for L times. We define the distribution of (z,a) For any h, define $\rho_h^k \in \Delta(\mathcal{Z} \times \mathcal{A})$ as follows:

$$\rho_h^k(z,a) = d_{\mathcal{P},h}^{\overline{\pi}_h^k \circ_L U(\mathcal{A})}(z,a).$$

Similarly, we define the distribution on $\mathcal{Z} \times \mathcal{A}$ after doing random action for $2L$ times. For any $k, h \geq 1$, we define β_h^k as follows:

$$\beta_h^k(z,a) = d_{\mathcal{P},h}^{\overline{\pi}_h^k \circ_{2L} U(\mathcal{A})}(z,a).$$

We also define the distribution induced by $\bar{\pi}^k$. For any k, h, we also define $\gamma_h^k \in \Delta(\mathcal{S} \times \mathcal{A})$ as follows:

$$\gamma_h^k(z,a) = d_{\mathcal{P},h}^{\overline{\pi}^k}(z,a).$$

For notational simplification, we denote $\|x\|_{\rho,\phi} = \|x\|_{\sum_{\rho,\phi}}$ and $\|x\|_{\rho^{-1},\phi} = \|x\|_{\sum_{\rho}^{-1},\phi}$ for $x \in \mathbb{R}^d$ and $\phi \in \Phi$, where $\sum_{\rho,\phi} = \mathbb{E}_{z,a\sim\rho}[\phi(z,a)\phi^\top(z,a)] + \lambda I$. We define $\|x\|_{\beta,\phi}$, $\|x\|_{\gamma,\phi}$ in the same way.

By Lemma 11, with probability at least $1-\delta$, we assume that we have learned a $\widehat{\phi}$ and $\widehat{\mu}$ such that

$$\mathbb{E}_{(z,a)\sim\rho_h}[\|\mathbb{P}_h^{\widehat{\mathcal{P}}}(\cdot \mid z,a)^\top - \mathbb{P}_h^{\mathcal{P}}(\cdot \mid z,a)^\top\|_1^2] \le \zeta_k, \forall h \in [H].$$
$$\mathbb{E}_{(z,a)\sim\beta_h}[\|\mathbb{P}_h^{\widehat{\mathcal{P}}}(\cdot \mid z,a)^\top - \mathbb{P}_h^{\mathcal{P}}(\cdot \mid z,a)^\top\|_1^2] \le \zeta_k, \forall h \in [H].$$

Here $\hat{\mathbb{P}}_h(o' \mid z,a) = \hat{\phi}(z,a)\hat{\mu}(o')$ and $\zeta_k = O(\log(|\mathcal{M}|)/K)$. In later arguments we will condition on these events. Next, we prove the almost optimism lemma restated below.

Lemma 18 (Almost Optimism). *Consider episode $k \in [K]$ and set*

$$\alpha_k = \sqrt{k|\mathcal{A}|^L\zeta_k + 4\lambda_k d}/c, \quad \lambda_k = O(d\log(|\mathcal{F}|k/\delta)).$$

where c is an absolute constant. Then with probability $1-\delta$,

$$V^{\pi^*,\widehat{\mathcal{P}}_k,r+\widehat{b}^k} - V^{\pi^*,\mathcal{P},r} \ge -\frac{c\alpha_k L}{\sqrt{k}}$$

holds for all $k \in [K]$.

Proof. By Lemma 17, we have

$$\begin{aligned}
&V^{\pi^*,\widehat{\mathcal{P}},r+\widehat{b}} - V^{\pi^*,\mathcal{P},r} \\
&= \sum_{h=0}^{H-1} \mathbb{E}_{(z_h,a_h)\sim d_{\widehat{\mathcal{P}},h}^{\pi^*}}[\widehat{b}_h(z_h,a_h) + \mathbb{E}_{o'\sim\mathbb{P}_h^{\widehat{\mathcal{P}}}(\cdot|z_h,a_h)}[V_{h+1}^{\pi^*,\mathcal{P},r}(z'_{h+1})] - \mathbb{E}_{o'\sim\mathbb{P}_h^{\mathcal{P}}(\cdot|z_h,a_h)}[V_{h+1}^{\pi^*,\mathcal{P},r}(z'_{h+1})]] \\
&\ge \sum_{h=0}^{H-1} \mathbb{E}_{(z_h,a_h)\sim d_{\widehat{\mathcal{P}},h}^{\pi^*}}\Bigg[\min(c\alpha_k\|\widehat{\phi}_h(z,a)\|_{\Sigma^{-1}_{\rho_h,\widehat{\phi}_h}}, 2) + \mathbb{E}_{o'\sim\mathbb{P}_h^{\widehat{\mathcal{P}}}(\cdot|z_h,a_h)}[V_{h+1}^{\pi^*,\mathcal{P},r}(z'_{h+1})] \\
&\qquad - \mathbb{E}_{o'\sim\mathbb{P}_h^{\mathcal{P}}(\cdot|z_h,a_h)}[V_{h+1}^{\pi^*,\mathcal{P},r}(z'_{h+1})]\Bigg],
\end{aligned} \tag{8}$$

where in the last step, we replace empirical covariance by population covariance by Lemma 16, here c is an absolute constant. Here $(z,a) \sim d_{\mathcal{P},h}^{\pi}$ means that (z,a) is sampled from transition $\mathcal{P}$ and policy π. We define

$$g_h(z,a) = \mathbb{E}_{o'_h\sim\mathbb{P}_h^{\widehat{\mathcal{P}}}(\cdot|z,a)}[V_{h+1}^{\pi^*,\mathcal{P},r}(c(z,a,o'_h))] - \mathbb{E}_{o'_h\sim\mathbb{P}_h^{\mathcal{P}}(\cdot|z,a)}[V_{h+1}^{\pi^*,\mathcal{P},r}(c(z,a,o'_h))].$$

Notice that we have $\|g_h\|_\infty \le 1$. By Lemma 11, for any (z,a) we have

$$\mathbb{E}_{(z,a)\sim\rho_h}[g_h^2(z,a)] \le \zeta_k,\ \mathbb{E}_{(z,a)\sim\beta_h}[g_h^2(z,a)] \le \zeta_k.$$

By Lemma 21, we have:

$$
\begin{aligned}
&\sum_{h=0}^{H-1} \mathbb{E}_{(z,a)\sim d^{\pi^*}_{\widehat{\mathcal{P}},h}}[g_h(z,a)] \\
\le &\sum_{h=1}^{H} \mathbb{E}_{(z_{h-L-1},a_{h-L-1})\sim d^{\pi}_{\widehat{\mathcal{P}},h-L-1}} \|\widehat{\phi}^\top(z_{h-L-1},a_{h-L-1})\|_{\rho^{-1}_{h-L-1},\widehat{\phi}_{h-L-1}} \\
&\qquad \cdot \sqrt{|A|^L k \cdot \mathbb{E}_{(\tilde{z}_h,\tilde{a}_h)\sim\beta_h}\{[g(\tilde{z}_h,\tilde{a}_h)]^2\} + B^2\lambda_k d} \\
\le &\sum_{h=1}^{H} \mathbb{E}_{(z_{h-L-1},a_{h-L-1})\sim d^{\pi}_{\widehat{\mathcal{P}},h-L-1}} \|\widehat{\phi}^\top(z_{h-L-1},a_{h-L-1})\|_{\rho^{-1}_{h-L-1},\widehat{\phi}_{h-L-1}} \\
&\qquad \cdot \sqrt{|A|^L k\zeta_k + B^2\lambda_k d} \\
= &c\alpha_k \sum_{h=1}^{H} \mathbb{E}_{(z_{h-L-1},a_{h-L-1})\sim d^{\pi}_{\widehat{\mathcal{P}},h-L-1}} \|\widehat{\phi}^\top(z_{h-L-1},a_{h-L-1})\|_{\rho^{-1}_{h-L-1},\widehat{\phi}_{h-L-1}},
\end{aligned}
\tag{9}
$$

where in the last step we define

$$
\alpha_k = \sqrt{k|\mathcal{A}|^L\zeta_k + 4\lambda_k d}/c.
$$

For $h \le 0$, we have

$$
\|\widehat{\phi}_h^\top(z_h,a_h)\|_{\rho_h^{-1},\widehat{\phi}_h} = \sqrt{\frac{1}{k+\lambda}} < \frac{1}{\sqrt{k}}. \tag{10}
$$

Combine (8), (9) and (10), we conclude the proof. □

With Lemma 18, we prove that under MDP $\mathcal{P}$ we can effectively learn the optimal policy with Algorithm 1.

Theorem 19. *With probability* $1-\delta$, *we have*

$$
\sum_{k=1}^{K} V^{\pi^*,\mathcal{P},r} - V^{\pi^k,\mathcal{P},r} \le O(H^2|\mathcal{A}|^L d^2 K^{1/2}\log(dK|\mathcal{F}|/\delta)^{1/2}).
$$

Proof. Similar to Lemma 18, we condition on the event that the MLE guarantee 11 holds, which happens with probability $1-\delta$. For fixed k we have

$$
\begin{aligned}
V^{\pi^*,\mathcal{P},r} - V^{\pi^k,\mathcal{P},r} &\le V^{\pi^*,\widehat{\mathcal{P}},r+\widehat{b}^k} - V^{\pi^k,\mathcal{P},r} + \frac{c\alpha_k L}{\sqrt{k}} \\
&\le V^{\pi^k,\widehat{\mathcal{P}},r+\widehat{b}^k} - V^{\pi^k,\mathcal{P},r} + \frac{c\alpha_k L}{\sqrt{k}},
\end{aligned}
$$

where the first inequality comes from Lemma 18, the second inequality comes from $\pi^k = \operatorname{argmax}_\pi V^{\pi,\widehat{\mathcal{P}},r+\widehat{b}^k}$. By Lemma 17, we have

$$
\begin{aligned}
&V^{\pi^k,\widehat{\mathcal{P}},r+\widehat{b}^k} - V^{\pi^k,\mathcal{P},r} + \frac{c\alpha_k L}{\sqrt{k}} \\
= &\sum_{h=0}^{H-1}\Big[\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}\big[\widehat{b}_h(z_h,a_h) + \mathbb{E}_{o'_h\sim\mathbb{P}_h^{\widehat{\mathcal{P}}}(o'_h|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{P}},r+\widehat{b}^k}(z'_{h+1})] \\
&\qquad - \mathbb{E}_{o'_h\sim\mathbb{P}_h^{\mathcal{P}}(o'_h|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{P}},r+\widehat{b}^k}(z'_{h+1})]\big]\Big] + \frac{c\alpha_k L}{\sqrt{k}},
\end{aligned}
$$

Further we have

$$V^{\pi^k,\widehat{\mathcal{P}},r+\widehat{b}^k} - V^{\pi^k,\mathcal{P},r} = \sum_{h=0}^{H-1}\Big[\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}\big[\widehat{b}_h(z_h,a_h) + \mathbb{E}_{o_h'\sim\mathbb{P}_h^{\widehat{\mathcal{P}}}(o_h'|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{P}},r+\widehat{b}^k}(z_{h+1}')] - \mathbb{E}_{o_h'\sim\mathbb{P}_h^{\mathcal{P}}(o_h'|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{P}},r+\widehat{b}^k}(z_{h+1}')]\big]\Big]$$

and the last equation comes from Lemma 17. Therefore we have

$$V^{\pi^*,\mathcal{P},r} - V^{\pi^k,\mathcal{P},r} \le \sum_{h=0}^{H-1}\Big[\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}\big[\widehat{b}_h(z_h,a_h) + \mathbb{E}_{o_h'\sim\mathbb{P}_h^{\widehat{\mathcal{P}}}(o_h'|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{P}},r+\widehat{b}^k}(z_{h+1}')] - \mathbb{E}_{o_h'\sim\mathbb{P}_h^{\mathcal{P}}(o_h'|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{P}},r+\widehat{b}^k}(z_{h+1}')]\big]\Big] + \frac{c\alpha_k L}{\sqrt{k}}$$

Denote

$$f_h(z_h,a_h) = \frac{1}{2H+1}\bigg(\mathbb{E}_{o_h'\sim\mathbb{P}_h^{\widehat{\mathcal{P}}}(o_h'|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{P}},r+\widehat{b}^k}(z_{h+1}')] - \mathbb{E}_{o_h'\sim\mathbb{P}_h^{\mathcal{P}}(o_h'|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{P}},r+\widehat{b}^k}(z_{h+1}')]\bigg).$$

Note that $\|\widehat{b}\|_\infty \le 2$, hence we have $\|V_{h+1}^{\pi^k,\widehat{\mathcal{P}},r+b}\|_\infty \le (2H+1)$. Combining this fact with the above expansion, we have

$$V^{\pi^*,\mathcal{P},r} - V^{\pi^k,\widehat{\mathcal{P}},r} = \sum_{h=0}^{H-1}\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}[\widehat{b}_h(z_h,a_h)] + (2H+1)\sum_{h=0}^{H-1}\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}[f_h(z_h,a_h)] + \frac{c\alpha_k L}{\sqrt{k}}. \tag{11}$$

First, we calculate the bonus term in (11). , we have

$$\begin{aligned}&\sum_{h=0}^{H-1}\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}[\widehat{b}_h(z_h,a_h)]\\ &\le \sum_{h=0}^{H-1}\mathbb{E}_{(\tilde{z},\tilde{a})\sim d_{h-L}^{\pi^k,\mathcal{P}}}\|\phi_{h-L}^*(\tilde{z},\tilde{a})\|_{\Sigma^{-1}_{\gamma_{h-L},\phi^*_{h-L}}}\sqrt{k|\mathcal{A}|^L\mathbb{E}_{(z,a)\sim\rho_h}[(\widehat{b}_h(z,a))^2] + 4\lambda_k d}.\end{aligned}$$

where the inequality is following Lemma 20 associate with $\|\widehat{b}_h\|_\infty \le 2$.

Note that we use the fact that $B = 2$ when applying Lemma 20. In addition, we have that for any $h \in [H]$,

$$k\mathbb{E}_{(z,a)\sim\rho_h}\Big[\|\widehat{\phi}_h(z,a)\|^2_{\Sigma^{-1}_{\rho_h,\widehat{\phi}_h}}\Big] = k\,\mathrm{Tr}\Big(\mathbb{E}_{\rho_h}[\widehat{\phi}_h\widehat{\phi}_h^\top]\{k\mathbb{E}_{\rho_h}[\widehat{\phi}_h\widehat{\phi}_h^\top] + \lambda_k I\}^{-1}\Big) \le d.$$

Then we have

$$\sum_{h=1}^{H}\mathbb{E}_{(z,a)\sim d_h^{\pi^k,\mathcal{P}}}[b_h(z,a)] \le \sum_{h=1}^{H}\mathbb{E}_{(\tilde{z},\tilde{a})\sim d_{h-L}^{\pi^k,\mathcal{P}}}\|\phi_{h-L}^*(\tilde{z},\tilde{a})\|_{\Sigma^{-1}_{\rho_{h-L},\phi^*_{h-L}}}\sqrt{|\mathcal{A}|^L\alpha_k^2 d + 4\lambda_k d}.$$

Now we bound the second term in (11). Further, with $\|f_h(z,a)\|_\infty \le 1$, we have

$$\begin{aligned}&\sum_{h=0}^{H-1}\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}[f_h(z_h,a_h)]\\ &\le \sum_{h=0}^{H-1}\mathbb{E}_{(\tilde{z},\tilde{a})\sim d_{h-L}^{\pi^k,\mathcal{P}}}\|\phi_{h-L}^*(\tilde{z},\tilde{a})\|_{\Sigma^{-1}_{\gamma_{h-L},\phi^*_{h-L}}}\sqrt{k|\mathcal{A}|^L\mathbb{E}_{(z,a)\sim\rho_h}[f_h^2(z,a)] + 4\lambda_k d}\\ &\le \sum_{h=0}^{H-1}\mathbb{E}_{(\tilde{z},\tilde{a})\sim d_{h-L}^{\pi^k,\mathcal{P}}}\|\phi_{h-L}^*(\tilde{z},\tilde{a})\|_{\Sigma^{-1}_{\gamma_{h-L},\phi^*_{h-L}}}\sqrt{k|\mathcal{A}|^L\zeta_k + 4\lambda_k d},\end{aligned}$$

where the first inequality is by By Lemma 20 and in the second inequality, we use $\mathbb{E}_{z,a\sim\rho_h}[f_h^2(z,a)] \le \zeta_k$. Then we have

$$\begin{aligned}
&V^{\pi^*,\mathcal{P},r} - V^{\pi^k,\mathcal{P},r} \\
&= \sum_{h=0}^{H-1} \mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}[b_h(z_h,a_h)] + (2H+1)\sum_{h=0}^{H-1} \mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}[f_h(z_h,a_h)] + \frac{c\alpha_k L}{\sqrt{k}} \\
&\le \sum_{h=0}^{H-1} \mathbb{E}_{(\tilde z,\tilde a)\sim d_{h-L}^{\pi^k,\mathcal{P}}} \|\phi_h^*(\tilde z,\tilde a)\|_{\Sigma^{-1}_{\gamma_{h-L},\phi^*_{h-L}}} \sqrt{|\mathcal{A}|^L\alpha_k^2 d + 4\lambda_k d} \\
&\quad + (2H+1)\sum_{h=0}^{H-1} \mathbb{E}_{(\tilde z,\tilde a)\sim d_{h-L}^{\pi^k,\mathcal{P}}} \|\phi_{h-L}^*(\tilde z,\tilde a)\|_{\Sigma_{\gamma^{-1}_{h-L},\phi^*_{h-L}}} \sqrt{k|\mathcal{A}|^L\zeta_k + 4\lambda_k d} + \frac{c\alpha_k L}{\sqrt{k}}.
\end{aligned}$$

Hereafter, we take the dominating term out. First, recall

$$\alpha_k = O(\sqrt{k|\mathcal{A}|^L\zeta_k + 4\lambda_k d}).$$

Second, recall that $\gamma_h^k(z,a) = \frac{1}{k}\sum_{i=0}^{k-1} d_h^{\pi^i}(z,a)$. Thus

$$\begin{aligned}
&\sum_{k=1}^{K} \mathbb{E}_{(\tilde z,\tilde a)\sim d_h^{\pi^k,\mathcal{P}}} \|\phi^*(\tilde z,\tilde a)\|_{\Sigma^{-1}_{\gamma_h^n,\phi_h^*}} \le \sqrt{K\sum_{k=1}^{K} \mathbb{E}_{(\tilde z,\tilde a)\sim d_h^{\pi^k,\mathcal{P}}}[\phi_h^*(\tilde z,\tilde a)^\top \Sigma^{-1}_{\gamma_h^k,\phi_h^*}\phi_h^*(\tilde z,\tilde a)]} \\
&\le \sqrt{K(\log\det(\sum_{k=1}^{K} \mathbb{E}_{(\tilde z,\tilde a)\sim d_h^{\pi^k,\mathcal{P}}}[\phi_h^*(\tilde z,\tilde a)\phi_h^*(\tilde z,\tilde a)^\top]) - \log\det(\lambda_1 I))} \\
&\le \sqrt{dK\log(1+\frac{K}{d\lambda_1}}),
\end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality, the second inequality is by Lemma 14 and the third inequality is by Lemma 15.

Finally, Lemma 11 gives

$$\zeta_k = O\left(\frac{\log(|\mathcal{F}|k/\delta)}{k}\right).$$

Combining all of the above, we have

$$\sum_{k=1}^{K} V^{\pi^*,\mathcal{P},r} - V^{\pi^k,\mathcal{P},r} \le O(H^2|\mathcal{A}|^L d^2 K^{1/2}\log(dK|\mathcal{F}|/\delta)^{1/2}),$$

which concludes the proof. ∎

Lemma 20 (L-step back inequality for the true model)**.** *Consider a set of functions $\{g_h\}_{h=0}^H$ that satisfies $g_h \in \mathcal{Z}\times\mathcal{A}\to\mathbb{R}$, s.t. $\|g_h\|_\infty \le B$ for all $h\in[H]$. Then, for any policy π, we have*

$$\begin{aligned}
\sum_{h=1}^{H} \mathbb{E}_\pi^{\mathcal{P}}[g(z_h,a_h)] \le \sum_{h=1}^{H} &\mathbb{E}^{\mathcal{P}}_{z_{h-L-1},a_{h-L-1}\sim\pi}\left[\|\phi(z_{h-L-1},a_{h-L-1})\|_{\beta^{-1}_{h-L-1},\phi_{h-L-1}}\right] \\
&\cdot\sqrt{|A|^L k\cdot\mathbb{E}_{(\tilde z_h,\tilde a_h)\sim\gamma_h}\{[g(\tilde z_h,\tilde a_h)]^2\} + B^2\lambda_k d}.
\end{aligned}$$

Proof. For $h\in[H]$ and $h'\in[h-L+1,h]$, we define $\mathcal{X}_l = \mathcal{S}^l\times\mathcal{O}^l\times\mathcal{A}^{l-1}$ and

$$x_{h'} = (s_{h-L+1:h'}, o_{h-L+1:h'}, a_{h-L+1:h'-1}),$$

where $l = h'-h+L-1$.

Now we define the moment matching policy $\mu^{\pi,h} = \{\mu_{h'}^{\pi,h} : \mathcal{X}_l \to \Delta(\mathcal{A})\}_{h'=h-L+1}^h$. We set $\mu^{\pi,h}$ as following:

$$\mu_{h'}^{\pi,h}(a_{h'} \mid x_{h'}) := \mathbb{E}_\pi^{\mathcal{P}}[\pi_{h'}(a_{h'} \mid z_{h'}) \mid x_{h'}] \text{ for } h'\le h-1, \text{ and } \mu_h^{\pi,h} = \pi_h. \tag{12}$$

Then we define policy $\tilde\pi_h$ which takes first $h-L$ actions from π and remaining actions from $\mu^{\pi,h}$.

Lemma B.2 in Efroni et al. (2022a): For a fixed $h \in [H]$, any $z_h \in \mathcal{Z}_h$ and fixed L-step policies π, $d_h^{\mathcal{P},\pi}(z_h) = d_h^{\mathcal{P},\tilde{\pi}^h}(z_h)$.

By Lemma B.2 in Efroni et al. (2022a), we have $d_h^{\mathcal{P},\pi}(z_h) = d_h^{\mathcal{P},\tilde{\pi}^h}(z_h)$. Then we have $d_h^{\mathcal{P},\pi}(z_h, a_h) = d_h^{\mathcal{P},\tilde{\pi}^h}(z_h, a_h)$ since $\mu_h^{\pi,h} = \pi_h$. Hence we have $\mathbb{E}_\pi^{\mathcal{P}}[g(z_h, a_h)] = \mathbb{E}_{\tilde{\pi}^h}^{\mathcal{P}}[g(z_h, a_h)]$.

Since $\mu^{\pi,h}$ is independent of s_{h-L+1}, we have the L-step-back decomposition:

$$
\begin{aligned}
&\mathbb{E}_{\tilde{\pi}^h}^{\mathcal{P}}[g(z_h, a_h)] \\
&= \mathbb{E}_{z_{h-L}, a_{h-L} \sim \pi}^{\mathcal{P}}\left[\int_{s_{h-L+1}} \phi^\top(z_{h-L}, a_{h-L})\omega(s_{h-L+1}) \cdot \mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}^{\mathcal{P}}[g(z_h, a_h) \mid s_{h-L+1}] \, \mathrm{d}s_{h-L+1}\right] \\
&= \mathbb{E}_{z_{h-L}, a_{h-L} \sim \pi}^{\mathcal{P}} \phi^\top(z_{h-L}, a_{h-L}) \cdot \int_{s_{h-L+1} \in \mathcal{S}} \omega(s_{h-L+1}) \cdot \mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}^{\mathcal{P}}[g(z_h, a_h) \mid s_{h-L+1}] \, \mathrm{d}s_{h-L+1} \\
&\le \mathbb{E}_{z_{h-L}, a_{h-L} \sim \pi}^{\mathcal{P}} \|\phi^\top(z_{h-L}, a_{h-L})\|_{\beta_{h-L}^{-1}, \phi_{h-L}} \\
&\qquad \cdot \left\| \int_{s_{h-L+1} \in \mathcal{S}} \omega(s_{h-L+1}) \cdot \mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}^{\mathcal{P}}[g(z_h, a_h) \mid s_{h-L+1}] \, \mathrm{d}s_{h-L+1} \right\|_{\beta_{h-L}, \phi_{h-L}}.
\end{aligned}
$$

The first equality comes from the definition of conditional expectation, and the inequality comes from Cauchy-Schwarz inequality. Here we use $\tilde{g}(s_{h-L})$ to denote $\mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}^{\mathcal{P}}[g(z_h, a_h) \mid s_{h-L}]$ for notational simplification.

We have

$$
\begin{aligned}
&\| \int_{s_{h-L+1} \in \mathcal{S}} \omega(s_{h-L+1})\tilde{g}(s_{h-L+1}) \, \mathrm{d}s_{h-L+1} \|_{\beta_{h-L}, \phi_{h-L}}^2 \\
&= \left\{ \int_{s_{h-L+1} \in \mathcal{S}} \omega(s_{h-L+1})\tilde{g}(s_{h-L+1}) \, \mathrm{d}s_{h-L+1} \right\}^\top \\
&\qquad \cdot \left\{ k\mathbb{E}_{(\tilde{z}_{h-L}, \tilde{a}_{h-L}) \sim \beta_{h-L+1}}[\phi_{h-L}(\tilde{z}_{h-L}, \tilde{a}_{h-L})\phi_{h-L}^\top(\tilde{z}_{h-L}, \tilde{a}_{h-L})] + \lambda_k I \right\} \\
&\qquad \cdot \left\{ \int_{s_{h-L+1} \in \mathcal{S}} \omega(s_{h-L+1})\tilde{g}(s_{h-L+1}) \, \mathrm{d}s_{h-L+1} \right\} \\
&\le k\mathbb{E}_{(\tilde{z}_{h-L}, \tilde{a}_{h-L}) \sim \beta_{h-L}} \left\{ [\int_{s_{h-L+1} \in \mathcal{S}} \omega^\top(s_{h-L+1})\phi_h(\tilde{z}_{h-L}, \tilde{a}_{h-L})\tilde{g}(s_{h-L+1}) \, \mathrm{d}s_{h-L+1}]^2 \right\} + B^2\lambda_k d \\
&= k\mathbb{E}_{(\tilde{z}_{h-L}, \tilde{a}_{h-L}) \sim \beta_{h-L}} \left\{ \left[\mathbb{E}_{s_{h-L+1} \sim \mathbb{P}_{h-L+1}^{\mathcal{P}}(\cdot \mid \tilde{z}_{h-L}, \tilde{a}_{h-L})} \mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}[g(z_h, a_h) \mid s_{h-L+1}] \right]^2 \right\} + B^2\lambda_k d,
\end{aligned}
$$

where the inequality comes from Cauchy-Schwarz inequality and $\|g_h\|_\infty \le B$. Moreover, we have

$$
\begin{aligned}
&k\mathbb{E}_{(\tilde{z}_{h-L}, \tilde{a}_{h-L}) \sim \beta_{h-L}} \left\{ [\mathbb{E}_{s_{h-L+1} \sim \mathbb{P}_{h-L+1}^{\mathcal{P}}(\cdot \mid \tilde{z}_{h-L}, \tilde{a}_{h-L})} \mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}[g(z_h, a_h) \mid s_{h-L+1}]]^2 \right\} \\
&\le k\mathbb{E}_{(\tilde{z}_{h-L}, \tilde{a}_{h-L}) \sim \beta_{h-L}, s_{h-L+1} \sim \mathbb{P}_{h-L+1}^{\mathcal{P}}(\cdot \mid \tilde{z}_{h-L}, \tilde{a}_{h-L}), a_{h-L+1:h} \sim \mu^{\pi,h}}[g(z_h, a_h)]^2 \\
&\le |A|^L k\mathbb{E}_{(\tilde{z}_{h-L}, \tilde{a}_{h-L}) \sim \beta_{h-L}, s_{h-L+1} \sim \mathbb{P}_{h-L+1}^{\mathcal{P}}(\cdot \mid \tilde{z}_{h-L}, \tilde{a}_{h-L}), a_{h-L+1:h} \sim U(\mathcal{A})}[g(z_h, a_h)]^2 \\
&= |A|^L k \cdot \mathbb{E}_{(\tilde{z}_h, \tilde{a}_h) \sim \gamma_h}\{[g(\tilde{z}_h, \tilde{a}_h)]^2\},
\end{aligned}
$$

where the first inequality is by Jensen's inequality and the second inequality is by importance sampling, the last equation is by the definition of γ_h.

Then, the final statement is immediately concluded. □

Lemma 21 (L-step back inequality for the learned model)**.** *Consider a set of functions $\{g_h\}_{h=0}^H$ that satisfies $g_h \in \mathcal{Z} \times \mathcal{A} \to$*

$\mathbb{R}$, *s.t.* $\|g_h\|_\infty \leq B$ *for all* $h \in [H]$. *We assume that MLE guarantee is hold. Then, for any policy* π*, we have*

$$\sum_{h=1}^{H} \mathbb{E}_\pi^{\widehat{\mathcal{P}}}[g(z_h, a_h)] \leq \sum_{h=1}^{H} \mathbb{E}_{z_{h-L-1}, a_{h-L-1} \sim \pi}^{\widehat{\mathcal{P}}} \left[\|\widehat{\phi}^\top (z_{h-L-1}, a_{h-L-1})\|_{\rho_{h-L-1}^{-1}, \widehat{\phi}_{h-L-1}} \right] \cdot \sqrt{|A|^L k \cdot \mathbb{E}_{(\tilde{z}_h, \tilde{a}_h) \sim \beta_h} \{[g(\tilde{z}_h, \tilde{a}_h)]^2\} + B^2 \lambda_k d} + kB^2 \zeta_k.$$

Proof. For $h \in [H]$ and $h' \in [h-L+1, h]$, we define $\mathcal{X}_l = \mathcal{S}^l \times \mathcal{O}^l \times \mathcal{A}^{l-1}$ and

$$x_{h'} = (s_{h-L+1:h'}, o_{h-L+1:h'}, a_{h-L+1:h'-1}),$$

where $l = h' - h + L - 1$.

Now we define the moment matching policy $\mu^{\pi,h} = \{\mu_{h'}^{\pi,h} : \mathcal{X}_l \to \Delta(\mathcal{A})\}_{h'=h-L+1}^{h}$. We set $\mu^{\pi,h}$ as following:

$$\mu_{h'}^{\pi,h}(a_{h'} \mid x_{h'}) := \mathbb{E}_\pi^{\widehat{\mathcal{P}}}[\pi_{h'}(a_{h'} \mid z_{h'}) \mid x_{h'}] \text{ for } h' \leq h-1, \text{ and } \mu_h^{\pi,h} = \pi_h.$$

Then we define policy $\tilde{\pi}_h$ which takes first $h - L$ actions from π and remaining actions from $\mu^{\pi,h}$.

By Lemma B.2 in Efroni et al. (2022a), we have $d_h^{\widehat{\mathcal{P}},\pi}(z_h) = d_h^{\widehat{\mathcal{P}},\tilde{\pi}^h}(z_h)$. Then we have $d_h^{\widehat{\mathcal{P}},\pi}(z_h, a_h) = d_h^{\widehat{\mathcal{P}},\tilde{\pi}^h}(z_h, a_h)$ since $\mu_h^{\pi,h} = \pi_h$. Hence we have $\mathbb{E}_\pi^{\widehat{\mathcal{P}}}[g(z_h, a_h)] = \mathbb{E}_{\tilde{\pi}^h}^{\widehat{\mathcal{P}}}[g(z_h, a_h)]$.

Since $\mu^{\pi,h}$ is independent of s_{h-L+1}, we have the L-step-back decomposition:

$$\begin{aligned}
&\mathbb{E}_{\tilde{\pi}^h}^{\widehat{\mathcal{P}}}[g(z_h, a_h)] \\
&= \mathbb{E}_{z_{h-L}, a_{h-L} \sim \pi}^{\widehat{\mathcal{P}}} \left[\int_{s_{h-L+1}} \widehat{\phi}^\top (z_{h-L}, a_{h-L}) \widehat{\omega}(s_{h-L+1})(s_{h-L+1}) \cdot \mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}^{\widehat{\mathcal{P}}}[g(z_h, a_h) \mid s_{h-L+1}] \, \mathrm{d}s_{h-L+1} \right] \\
&= \mathbb{E}_{z_{h-L}, a_{h-L} \sim \pi}^{\widehat{\mathcal{P}}} \widehat{\phi}^\top (z_{h-L}, a_{h-L}) \cdot \int_{s_{h-L+1} \in \mathcal{S}} \widehat{\omega}(s_{h-L+1}) \cdot \mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}^{\widehat{\mathcal{P}}}[g(z_h, a_h) \mid s_{h-L+1}] \, \mathrm{d}s_{h-L+1} \\
&\leq \mathbb{E}_{z_{h-L}, a_{h-L} \sim \pi}^{\widehat{\mathcal{P}}} \|\widehat{\phi}^\top (z_{h-L}, a_{h-L})\|_{\rho_{h-L}^{-1}, \widehat{\phi}_{h-L}} \\
&\qquad \cdot \left\| \int_{s_{h-L+1} \in \mathcal{S}} \widehat{\omega}(s_{h-L+1}) \cdot \mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}^{\widehat{\mathcal{P}}}[g(z_h, a_h) \mid s_{h-L+1}] \, \mathrm{d}s_{h-L+1} \right\|_{\rho_{h-L}, \widehat{\phi}_{h-L}},
\end{aligned}$$

where the inequality is by Cauchy-Schwarz inequality.

Now we use $\tilde{g}(s_{h-L})$ to denote $\mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}^{\widehat{\mathcal{P}}}[g(z_h, a_h) \mid s_{h-L}]$ for notational simplification. We have

$$\begin{aligned}
&\| \int_{s_{h-L+1} \in \mathcal{S}} \widehat{\omega}(s_{h-L+1}) \tilde{g}(s_{h-L+1}) \, \mathrm{d}s_{h-L+1} \|_{\rho_{h-L}, \widehat{\phi}_{h-L}}^2 \\
&= \left\{ \int_{s_{h-L+1} \in \mathcal{S}} \widehat{\omega}(s_{h-L+1}) \tilde{g}(s_{h-L+1}) \, \mathrm{d}s_{h-L+1} \right\}^\top \\
&\qquad \left\{ k \mathbb{E}_{(\tilde{z}_{h-L}, \tilde{a}_{h-L}) \sim \rho_{h-L+1}} [\widehat{\phi}_{h-L+1}(\tilde{z}_{h-L}, \tilde{a}_{h-L}) \widehat{\phi}_{h-L+1}^\top (\tilde{z}_{h-L}, \tilde{a}_{h-L})] + \lambda_k I \right\} \\
&\qquad \cdot \left\{ \int_{s_{h-L+1} \in \mathcal{S}} \widehat{\omega}(s_{h-L+1}) \tilde{g}(s_{h-L+1}) \, \mathrm{d}s_{h-L+1} \right\} \\
&\leq k \mathbb{E}_{(\tilde{z}_{h-L}, \tilde{a}_{h-L}) \sim \rho_{h-L}} \left\{ \left[\int_{s_{h-L+1} \in \mathcal{S}} \widehat{\omega}^\top (s_{h-L+1}) \widehat{\phi}_{h-L+1}(\tilde{z}_{h-L}, \tilde{a}_{h-L}) \tilde{g}(s_{h-L+1}) \, \mathrm{d}s_{h-L+1} \right]^2 \right\} + B^2 \lambda_k d \\
&= k \mathbb{E}_{(\tilde{z}_{h-L}, \tilde{a}_{h-L}) \sim \rho_{h-L}} \left\{ \left[\mathbb{E}_{s_{h-L+1} \sim \widehat{\mathcal{P}}_{h-L+1}(\tilde{z}_{h-L}, \tilde{a}_{h-L})} \mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}[g(z_h, a_h) \mid s_{h-L+1}] \right]^2 \right\} + B^2 \lambda_k d,
\end{aligned}$$

where inequality is because $\|g_h\| \leq B$. Moreover, we have

$$
\begin{aligned}
& k\mathbb{E}_{(\tilde{z}_{h-L},\tilde{a}_{h-L})\sim\rho_{h-L}}\left\{[\mathbb{E}_{s_{h-L+1}\sim\widehat{\mathcal{P}}_{h-L+1}(\tilde{z}_{h-L},\tilde{a}_{h-L})}\mathbb{E}_{a_{h-L+1:h}\sim\mu^{\pi,h}}[g(z_h,a_h)\mid s_{h-L+1}]]^2\right\} \\
& \leq k\mathbb{E}_{(\tilde{z}_{h-L},\tilde{a}_{h-L})\sim\rho_{h-L}}\left\{[\mathbb{E}_{s_{h-L+1}\sim\mathbb{P}^{\mathcal{P}}_{h-L+1}(\cdot|\tilde{z}_{h-L},\tilde{a}_{h-L})}\mathbb{E}_{a_{h-L+1:h}\sim\mu^{\pi,h}}[g(z_h,a_h)\mid s_{h-L+1}]]^2\right\}+kB^2\zeta_k \\
& \leq k\mathbb{E}_{(\tilde{z}_{h-L},\tilde{a}_{h-L})\sim\rho_{h-L},s_{h-L+1}\sim\mathbb{P}^{\mathcal{P}}_{h-L+1}(\cdot|\tilde{z}_{h-L},\tilde{a}_{h-L}),a_{h-L+1:h}\sim\mu^{\pi,h}}[g(z_h,a_h)\mid s_{h-L+1}]^2+kB^2\zeta_k \\
& \leq |A|^L k\mathbb{E}_{(\tilde{z}_{h-L},\tilde{a}_{h-L})\sim\rho_{h-L},s_{h-L+1}\sim\mathbb{P}^{\mathcal{P}}_{h-L+1}(\cdot|\tilde{z}_{h-L},\tilde{a}_{h-L}),a_{h-L+1:h}\sim U(\mathcal{A})}[g(z_h,a_h)]^2+kB^2\zeta_k \\
& = |A|^L k\cdot\mathbb{E}_{(\tilde{z}_h,\tilde{a}_h)\sim\beta_h}\{[g(\tilde{z}_h,\tilde{a}_h)]^2\}+kB^2\zeta_k,
\end{aligned}
$$

where the first inequality is by MLE guarantee, the second inequality is by Jensen's inequality and the last inequality is by importance sampling.

Then, the final statement is immediately concluded. □

C. MLE guarantee

In this section, we present the MLE guarantee used for L-step decodable POMDPs and γ-observable POMDPs. Regarding the proof, refer to Agarwal et al. (2020).

Lemma 22 (MLE guarantee)**.** *Set $\lambda_k = \Theta(d\log(|\mathcal{F}|k/\delta))$, for any time step $h \in [H]$, denote ρ_h as the joint distribution for (z, a) in the dataset $\mathcal{D}$ of step k, with probability at least $1-\delta$ we have*

$$
\mathbb{E}_{z,a\sim\rho}[\|\mathbb{P}^{\widehat{\mathcal{P}}}_h(\cdot\mid z,a)^\top - \mathbb{P}^{\mathcal{P}}_h(\cdot\mid z,a)^\top\|_1^2] \leq \zeta_k, \tag{13}
$$

recall that $\zeta_k = O(\log(|\mathcal{F}|k/\delta)/k)$.

Lemma 23 (MLE guarantee for POMDP)**.** *Consider parameters defined in Theorem 10 and time step h. Denote ρ_h as the joint distribution for $(o_{3-2L:h}, a_{3-2L:h}, o'_{h+1})$ in the dataset $\mathcal{D}$ of size k. Then, with probability at least $1-\delta$ we have*

$$
\mathbb{E}_{o_{3-2L:h},a_{3-2L:h}\sim\rho}[\|\mathbb{P}^{\widehat{\mathcal{P}}}_h(\cdot\mid o_{3-2L:h},a_{3-2L:h})^\top - \mathbb{P}^{\mathcal{P}}_h(\cdot\mid o_{3-2L:h},a_{3-2L:h})^\top\|_1^2] \leq \zeta_k = O\left(\frac{\log(k|\mathcal{F}|/\delta)}{k}\right),
$$

in addition, we have

$$
\mathbb{E}_{z_h,a_h\sim\rho}[\|\mathbb{P}^{\widehat{\mathcal{M}}}_h(\cdot\mid z_h,a_h)^\top - \mathbb{P}^{\mathcal{M}}_h(\cdot\mid z_h,a_h)^\top\|_1^2] \leq O\left(\frac{\log(k|\mathcal{F}|/\delta)}{k}+\epsilon_1\right).
$$

D. Missing Proofs of Section 5.1

D.1. Construction of the approximated MDP

To construct this approximated MDP, we need to first calculate the belief state and approximated belief state, which is the conditional probability of state s_h given the true transition and an action and observation sequence $\{o_{3-2L}, a_{3-2L}, \cdots, a_h, o_h\}$ and $1 \leq h \leq H$ respectively.

Consider a POMDP and a history $(o_{3-2L:h}, a_{3-2L:h-1}) \in \mathcal{H}_h$, the belief state $\mathbf{b}^{\mathcal{P}}_h(o_{3-2L:h}, a_{3-2L:h-1}) \in \Delta(\mathcal{S})$ is given by the distribution of the state s_h conditioned on taking actions $a_{1:h-1}$ and observing $o_{1:h}$ in the first h steps. Formally, the belief state is defined inductively as follows: $\mathbf{b}^{\mathcal{P}}_1(\varnothing) = b_1$, where b_1 is a properly chosen prior distribution whose precise form is deferred to appendix C. For $2 \leq h \leq H$ and any $(o_{3-2L:h}, a_{3-2L:h-1}) \in \mathcal{H}_h$, define

$$
\mathbf{b}^{\mathcal{P}}_h(o_{3-2L:h}, a_{3-2L:h-1}) := U^{\mathcal{P}}_{h-1}(\mathbf{b}^{\mathcal{P}}_{h-1}(a_{1:h-2}, o_{2:h-1}); a_{h-1}, o_h),
$$

where for $\mathbf{b} \in \Delta(\mathcal{S}), a \in \mathcal{A}, o \in \mathcal{O}$, the belief update operator $U^{\mathcal{P}}_h$ is defined as

$$
U^{\mathcal{P}}_h(\mathbf{b}; a, o)(s) := \frac{\mathbb{O}_{h+1}(o\mid s)\cdot\sum_{s'\in\mathcal{S}}\mathbf{b}(s')\cdot\mathbb{P}_h(s\mid s',a)}{\sum_{x\in\mathcal{S}}\mathbb{O}_{h+1}(o\mid x)\sum_{s'\in\mathcal{S}}\mathbf{b}(s')\cdot\mathbb{P}_h(x\mid s',a)},
$$

This operator calculates the belief state for $h+1$-step when the belief for the h-step is b, and after the agent takes action a and receives the observation o.

Recall that $\mathcal{P}$ has a low-rank structure. For an extended POMDP, we have

$$\mathbb{P}_h^{\mathcal{P}}(o_{h+1} \mid o_{3-2L:h}, a_{3-2L:h}) = \int_{\mathcal{S}_{h+1}} \omega_h(s_{h+1}) \cdot \mathbb{O}_{h+1}(o_{h+1} \mid s') \, \mathrm{d}s' \cdot \int_{\mathcal{S}} \psi_h(s, a_h) b_h^{\mathcal{P}}(o_{3-2L:h}, a_{3-2L:h-1})(s) \, \mathrm{d}s \,. \tag{14}$$

For $h \in [H]$, $\tau_h \in \mathcal{H}_h$, $o_h \in \mathcal{O}$ and $a_h \in \mathcal{Z}_h$, we denote

$$\int \psi_h(s_h, a_h) b_h^{\mathcal{P}}(\tau_h)(s_h) \, \mathrm{d}s_h = \xi_h(\tau_h, a_h), \tag{15}$$

we define the approximated belief $\bar{b}_h(o_{h-L:h}, a_{h-L:h-1})$ to approximate the true belief $b_h(o_{3-2L:h}, a_{3-2L:h-1})$.

For $\mathbf{b} \in \Delta(\mathcal{S})$ and $o \in \mathcal{O}$, we define $B(\mathbf{b}, o)$ as the operation that incorporates observation o by

$$B(\mathbf{b}, o)(s) := \frac{\mathbb{O}(o \mid s) \cdot \mathbf{b}(s)}{\sum_{x \in \mathcal{S}} \mathbb{O}(o \mid x) \sum_{s' \in \mathcal{S}} \mathbf{b}(s')},$$

which denotes the belief distribution after receiving the observation o as the original belief distribution was b.

For an action and observation sequence $\{o_{3-2L}, a_{3-2L}, \cdots, o_H, a_H\}$ and $2 \le h \le H$. We define the approximated belief as :

$$\begin{aligned}
&\bar{b}_{h-L} = B(\tilde{b}_0^{h-L}, o_{h-L}), \qquad (16)\\
&\bar{b}_{h-L+\tau}(o_{h-L:h-L+\tau}, a_{h-L:h-L-1+\tau}) = U_{h-L-1+\tau}^{\mathcal{P}}(\bar{b}_{h-L-1+\tau}(o_{h-L:h-L-1+\tau}, a_{h-L:\tau}), o_{h-L+\tau}, a_{h-L-1+\tau}), 1 \le \tau \le l,
\end{aligned}$$

where the construction of the initial belief $\tilde{b}_0^{h-L}$ can be found in Appendix D.2.

We have the following lemma to give an upper bound for the difference between the approximate belief and the true belief. The detailed proof can be found in Appendix D.3.

Lemma 24. *For any policy $\pi \in \Delta(\mathcal{A})$ and $h \ge 2-L$, the approximate belief $\bar{b}$ defined in (16) satisfy that*

$$\mathbb{E}^{\pi, \mathcal{P}}[\|b_{h+L}(o_{3-2L:h+L}, a_{3-2L:h+L-1}) - \bar{b}_{h+L}(z_h, a_h)\|_1] \le \epsilon_1, \text{ where } (z_h, a_h) = (o_{h:h+L}, a_{h:h+L-1}).$$

We have

$$\int \omega_h(s_h') \cdot \mathbb{O}_h(o_h \mid s_h') \, \mathrm{d}s_h' = \mu_h(o_h) \tag{17}$$

Now we can construct $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r)$ as following

$$\begin{aligned}
&\mathbb{P}_h^{\mathcal{M}}(o_{h+1} \mid o_{h+1-L:h}, a_{h+1-L:h}) \\
&= \int_{\mathcal{S}_{h+1}} \omega_{h+1}(s_{h+1}) \cdot \mathbb{O}_{h+1}(o_{h+1} \mid s_{h+1}) \, \mathrm{d}s_{h+1} \cdot \int_{\mathcal{S}_h} \psi(s_h, a_h) \bar{b}_h^{\mathcal{P}}(o_{h-L+1:h}, a_{h-L+1:h-1})(s_h) \, \mathrm{d}s_h \,.
\end{aligned}$$

We denote

$$\int \psi(s_h, a_h) \bar{b}_h^{\mathcal{P}}(z_h)(s_h) \, \mathrm{d}s_h = \phi(z_h, a_h). \tag{18}$$

We define $(\mu, \phi, \xi) = q(\mathbb{O}, \omega, \psi)$ by (15), (17) and (18).

D.2. Construction of $\tilde{b}_0^h$

Now we present the construction of $\tilde{b}_0^h$ for $h \in [H]$, which is done by Uehara et al. (2022b).

Lemma 25 (G-optimal design (Uehara et al., 2022b)). *Suppose $\mathcal{X} \in \mathbb{R}^d$ is a compact set. There exists a distribution ρ over $\mathcal{X}$ such that: (i) ρ is supported on at most $d(d+1)/2$ points. (ii) For any $x' \in \mathcal{X}$, we have $x'^\top \mathbb{E}_{x\sim\rho}\left[xx^\top\right]^{-1} x' \le d$.*

By Lemma 25, there exist $\rho_h \in \Delta(\mathcal{S}_h \times \mathcal{A}_h)$ to be the G-optimal design for $\psi_h(s,a)$. Denote the support of ρ_h to be $S_{\rho,h}$, $S_{\rho,h}$ has most $d(d+1)/2$ points. We denote $S_{\rho,h} = \{s_h^i, a_h^i\}_{i=1}^{|S_{\rho,h}|}$.

We set $\tilde{b}_0^h$ as

$$\tilde{b}_0^h(\cdot) := \sum_{\tilde{s},\tilde{a}} \rho_h(\tilde{s},\tilde{a}) \mathbb{P}_h^{\mathcal{P}}(\cdot \mid \tilde{s},\tilde{a}) = \sum_{i=1}^{|S_{\rho,h}|} \rho\left(s_h^i, a_h^i\right) \mathbb{P}_h^{\mathcal{P}}\left(\cdot \mid s_h^i, a_h^i\right)$$

D.3. Proof for Proposition 8

First, we prove the existence of the approximated MDP $\mathcal{M}$ in Proposition 8.

Proof of Lemma 24. The proof is same to Theorem 14 of Uehara et al. (2022a), the only difference is that the process is start from $3-2L$ instead of 1. $\square$

Now we construct the approximated MDP $\mathcal{M}$. For a state $z_h = (o_{h-L:h}, a_{h-L:h-1}) \in \mathcal{Z}$, action a_h, and subsequent observation $o_{h+1} \in \mathcal{O}$, define

$$\mathbb{P}_h^{\mathcal{M}}((o_{h-L+1:h+1}, a_{h-L+1:h}) \mid z_h, a_h) := e_{o_{h+1}}^\top \cdot \mathbb{O}_{h+1}^{\mathcal{P}} \cdot \mathbb{T}_h^{\mathcal{P}}(a_h) \cdot \bar{b}_h(o_{h-L:h}, a_{h-L:h-1}).$$

Hence we can define

$$\phi(z_h, a_h) = \mathbb{T}_h^{\mathcal{P}}(a_h) \cdot \bar{b}_h,\ \mu(o_{h+1}) = e_{o_{h+1}}^\top \cdot \mathbb{O}_{h+1}^{\mathcal{P}}$$

Note that for the POMDP $\mathcal{P}$, we have

$$\mathbb{P}_{o_{h+1}\sim\pi}^{\mathcal{P}}(o_{h+1} \mid o_{1:h}, a_{1:h}) = e_{o_{h+1}}^\top \cdot \mathbb{O}_{h+1}^{\mathcal{P}} \cdot \mathbb{T}_h^{\mathcal{P}}(a_h) \cdot \mathbf{b}_h^{\mathcal{P}}(o_{1:h-1}, a_{1:h}).$$

Lemma 24 shows that $\mathbb{E}[\|b_{h+L}(o_{3-2L:h+L}, a_{3-2L:h+L-1}) - \bar{b}_{h+L}(o_{h:h+L}, a_{h:h+L-1})\|_1; a_{1:h+L-1} \sim \pi]$ is small in expectation under any policy π, so we have

$$\mathbb{E}_{a_{1:h}, o_{2:h}\sim\pi} \|\mathbb{P}_h^{\mathcal{M}}(o_{h+1} \mid z_h, a_h) - \mathbb{P}_h^{\mathcal{P}}(o_{h+1} \mid o_{1:h}, a_{1:h})\|_1 \le \epsilon_1.$$

With the approximated MDP and approximated belief, we have the following lemma.

Lemma 26. *For any function $g : \mathcal{S} \to \mathbb{R}$, $\|g\|_\infty \le 1$, we have*

$$\mathbb{E}_{a_{1:h}, o_{2:h}\sim\pi} \int_{s_{h+1}} \mathbb{P}_h^{\mathcal{P}}(s_{h+1} \mid o_{1:h}, a_{1:h}) g(s_{h+1}) \,\mathrm{d}s_{h+1} \le \mathbb{E}_{a_{1:h}, o_{2:h}\sim\pi} \int_{s_{h+1}} \omega(s_{h+1})^\top \phi(z_h, a_h) g(s_{h+1}) \,\mathrm{d}s_{h+1} + \epsilon_1.$$

Proof. We have

$$\begin{aligned}
&\mathbb{E}_{a_{1:h}, o_{2:h}\sim\pi} \int_{s_{h+1}} \mathbb{P}_h^{\mathcal{P}}(s_{h+1} \mid o_{1:h}, a_{1:h}) g(s_{h+1}) \,\mathrm{d}s_{h+1} \\
&\quad = \mathbb{E}_{a_{1:h}, o_{2:h}\sim\pi} \int_{s_{h+1}} \omega(s_{h+1})^\top \int_{s_h} \psi(s_h, a_h) b_h^{\mathcal{P}}(o_{3-2L:h}, a_{3-2L:h})(s_h) g(s_{h+1}) \,\mathrm{d}s_h \,\mathrm{d}s_{h+1} \\
&\quad \le \mathbb{E}_{a_{1:h}, o_{2:h}\sim\pi} \int_{s_{h+1}} \omega(s_{h+1})^\top \int_{s_h} \psi(s_h, a_h) |(b_h^{\mathcal{P}}(o_{3-2L:h}, a_{3-2L:h}) - \bar{b}_h(z_h, a_h))(s_h)|) |g(s_{h+1})| \,\mathrm{d}s_h \,\mathrm{d}s_{h+1} \\
&\qquad + \mathbb{E}_{a_{1:h}, o_{2:h}\sim\pi} \int_{s_{h+1}} \omega(s_{h+1})^\top \int_{s_h} \psi(s_h, a_h) \bar{b}_h(z_h, a_h)(s_h) g(s_{h+1}) \,\mathrm{d}s_h \,\mathrm{d}s_{h+1},
\end{aligned}$$

where the inequality is by Cauchy-Schwarz inequality.

Since we have $\int_{s_{h+1}} \omega(s_{h+1})^\top \psi(s_h, a_h)|g(s_{h+1})| \, \mathrm{d}s_{h+1} \leq 1$ for any s_h, so we have

$$\begin{aligned}
&\mathbb{E}_{a_{1:h}, o_{2:h} \sim \pi} \int_{s_{h+1}} \omega(s_{h+1})^\top \int_{s_h} \psi(s_h, a_h) |(b_h^{\mathcal{P}}(o_{3-2L:h}, a_{3-2L:h}) - \bar{b}_h(z_h, a_h))(s_h)||g(s_{h+1})| \, \mathrm{d}s_h \, \mathrm{d}s_{h+1} \\
&\leq \mathbb{E}_{a_{1:h}, o_{2:h} \sim \pi} \int_{s_h} \left|(b_h^{\mathcal{P}}(o_{3-2L:h}, a_{3-2L:h}) - \bar{b}_h(z_h, a_h))(s_h)\right| \mathrm{d}s_h \\
&\leq \epsilon_1,
\end{aligned}$$

where the last inequality is by Lemma 24.

Remember that $\int_{\mathcal{S}} \psi(s, a_{h-L}) \bar{b}_h(z_{h-L}, a_{h-L})(s) \, \mathrm{d}s = \phi_h(z_h, a_h)$, so we have

$$\begin{aligned}
&\mathbb{E}_{a_{1:h}, o_{2:h} \sim \pi} \int_{s_{h+1}} \mathbb{P}_h^{\mathcal{P}}(s_{h+1} \mid o_{1:h}, a_{1:h}) g(s_{h+1}) \, \mathrm{d}s_{h+1} \\
&\quad \leq \mathbb{E}_{a_{1:h}, o_{2:h} \sim \pi} \int_{s_{h+1}} \omega(s_{h+1})^\top \phi(z_h, a_h) g(s_{h+1}) \, \mathrm{d}s_{h+1} + \epsilon_1,
\end{aligned}$$

which concludes the proof. □

D.4. Proof of Lemma 9

Proof of Lemma 9.

$$\begin{aligned}
|V_1^{\pi, \mathcal{P}, r}(\varnothing) - V_1^{\pi, \mathcal{M}, r}(\varnothing)| &= \left| \mathbb{E}^{\mathcal{P}}_{(a_{1:H-1}, o_{2:H}) \sim \pi} \left[\sum_{h=1}^{H} ((\mathbb{P}_h^{\mathcal{P}} - \mathbb{P}_h^{\mathcal{M}})(V_{h+1}^{\pi, \mathcal{M}, r} + r_{h+1}))(a_{1:h}, o_{2:h}) \right] \right| \\
&= \left| \mathbb{E}^{\mathcal{P}}_{(a_{1:H-1}, o_{2:H}) \sim \pi} \left[\sum_{h=1}^{H-1} ((\mathbb{P}_h^{\mathcal{P}} - \mathbb{P}_h^{\mathcal{M}})(V_{h+1}^{\pi, \mathcal{M}, r} + r_{h+1}))(a_{1:h}, o_{2:h}) \right] \right| \\
&\leq \frac{H}{2} \cdot \mathbb{E}^{\mathcal{P}}_{(o_{3-2L:h}, a_{3-2L:h-1}) \sim \pi} \left[\sum_{h=1}^{H-1} \|\mathbb{P}_h^{\mathcal{P}}(\cdot \mid a_{1:h}, o_{2:h}) - \mathbb{P}_h^{\mathcal{M}}(\cdot \mid z_h, a_h)\|_1 \right] \\
&\leq \frac{H^2 \epsilon_1}{2},
\end{aligned}$$

where the first inequality uses the fact that $|(V_{h+1}^{\pi, \mathcal{M}, r} + r_{h+1})(a_{1:h}, o_{2:h+1})| \leq H$ for all $a_{1:h}, o_{2:h+1}$, the second inequality is by Proposition 8. □

Lemma 27 (L-step back inequality for the true POMDP)**.** *Consider a set of functions $\{g_h\}_{h=0}^H$ that satisfies $g_h \in \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$, s.t. $\|g_h\|_\infty \leq B$ for all $h \in [H]$. Then, for any policy π, we have*

$$\begin{aligned}
\sum_{h=1}^{H} \mathbb{E}_\pi^{\mathcal{P}}[g(z_h, a_h)] \leq \sum_{h=1}^{H} &\mathbb{E}^{\mathcal{P}}_{z_{h-L-1}, a_{h-L-1} \sim \pi} \|\phi^\top(z_{h-L-1}, a_{h-L-1})\|_{\beta_{h-L-1}^{-1}, \phi_{h-L-1}} \\
&\cdot \sqrt{|A|^L k \cdot \mathbb{E}^{\mathcal{P}}_{(\tilde{z}_h, \tilde{a}_h) \sim \gamma_h} \{[g(\tilde{z}_h, \tilde{a}_h)]^2\} + B^2 \lambda_k d + k B^2 \epsilon_1} + B \epsilon_1.
\end{aligned}$$

Proof. For $h \in [H]$ and $h' \in [h - L + 1, h]$, we define $\mathcal{X}_l = \mathcal{S}^l \times \mathcal{O}^l \times \mathcal{A}^{l-1}$ and

$$x_{h'} = (s_{h-L+1:h'}, o_{h-L+1:h'}, a_{h-L+1:h'-1}),$$

where $l = h' - h + L - 1$.

Now we define the moment matching policy $\mu^{\pi, h} = \{\mu_{h'}^{\pi, h} : \mathcal{X}_l \to \Delta(\mathcal{A})\}_{h'=h-L+1}^{h}$. We set $\mu^{\pi, h}$ as following:

$$\mu_{h'}^{\pi, h}(a_{h'} \mid x_{h'}) := \mathbb{E}_\pi^{\mathcal{P}}[\pi_{h'}(a_{h'} \mid z_{h'}) \mid x_{h'}] \text{ for } h' \leq h - 1, \text{ and } \mu_h^{\pi, h} = \pi_h.$$

Then we define policy $\tilde{\pi}^h$ which takes first $h-L$ actions from π and remaining actions from $\mu^{\pi,h}$.

By Lemma B.2 in Efroni et al. (2022a), we have $d_h^{\mathcal{P},\pi}(z_h) = d_h^{\mathcal{P},\tilde{\pi}^h}(z_h)$. Then we have $d_h^{\mathcal{P},\pi}(z_h,a_h) = d_h^{\mathcal{P},\tilde{\pi}^h}(z_h,a_h)$ since $\mu_h^{\pi,h} = \pi_h$. Hence we have $\mathbb{E}_\pi^{\mathcal{P}}[g(z_h,a_h)] = \mathbb{E}_{\tilde{\pi}^h}^{\mathcal{P}}[g(z_h,a_h)]$.

Since $\mu^{\pi,h}$ is independent of s_{h-L+1}, we have the L-step-back decomposition:

$$
\begin{aligned}
&\mathbb{E}_{\tilde{\pi}^h}^{\mathcal{P}}[g(z_h,a_h)] \\
&= \mathbb{E}^{\mathcal{P}}_{o_{3-2L:h-L},a_{3-2L,h-L}\sim\pi}\left[\int_{s_{h-L+1}} \mathbb{P}(s_{h-L+1}\mid o_{3-2L:h-L},a_{3-2L,h-L})\cdot \mathbb{E}^{\mathcal{P}}_{a_{h-L+1:h}\sim\mu^{\pi,h}}[g(z_h,a_h)\mid s_{h-L+1}]\,\mathrm{d}s_{h-L+1}\right] \\
&\le \mathbb{E}^{\mathcal{P}}_{z_{h-L},a_{h-L}\sim\pi}\left[\int_{s_{h-L+1}} \phi^\top(z_{h-L},a_{h-L})\omega(s_{h-L+1})\cdot \mathbb{E}^{\mathcal{P}}_{a_{h-L+1:h}\sim\mu^{\pi,h}}[g(z_h,a_h)\mid s_{h-L+1}]\,\mathrm{d}s_{h-L+1}\right] + B\epsilon_1 \\
&= \mathbb{E}^{\mathcal{P}}_{z_{h-L},a_{h-L}\sim\pi}\phi^\top(z_{h-L},a_{h-L})\cdot\int_{s_{h-L+1}\in\mathcal{S}}\omega(s_{h-L+1})\cdot \mathbb{E}^{\mathcal{P}}_{a_{h-L+1:h}\sim\mu^{\pi,h}}[g(z_h,a_h)\mid s_{h-L+1}]\,\mathrm{d}s_{h-L+1} + B\epsilon_1 \\
&\le \mathbb{E}^{\mathcal{P}}_{z_{h-L},a_{h-L}\sim\pi}\|\phi^\top(z_{h-L},a_{h-L})\|_{\beta_{h-L}^{-1},\phi_{h-L}} \\
&\qquad \cdot \left\|\int_{s_{h-L+1}\in\mathcal{S}}\omega(s_{h-L+1})\cdot \mathbb{E}^{\mathcal{P}}_{a_{h-L+1:h}\sim\mu^{\pi,h}}[g(z_h,a_h)\mid s_{h-L+1}]\,\mathrm{d}s_{h-L+1}\right\|_{\beta_{h-L},\phi_{h-L}} + B\epsilon_1,
\end{aligned}
$$

where the first inequality is because Lemma 26.

Now we use $\tilde{g}(s_{h-L})$ to denote $\mathbb{E}^{\mathcal{P}}_{a_{h-L+1:h}\sim\mu^{\pi,h}}[g(z_h,a_h)\mid s_{h-L}]$ for notational simplification. We have

$$
\begin{aligned}
&\left\|\int_{s_{h-L+1}\in\mathcal{S}}\omega(s_{h-L+1})\tilde{g}(s_{h-L+1})\,\mathrm{d}s_{h-L+1}\right\|^2_{\beta_{h-L},\phi_{h-L}} \\
&= \left\{\int_{s_{h-L+1}\in\mathcal{S}}\omega(s_{h-L+1})\tilde{g}(s_{h-L+1})\,\mathrm{d}s_{h-L+1}\right\}^\top \\
&\qquad \cdot\left\{k\mathbb{E}_{(\tilde{z}_{h-L},\tilde{a}_{h-L})\sim\beta_{h-L+1}}[\phi_{h-L}(\tilde{z}_{h-L},\tilde{a}_{h-L})\phi_{h-L}^\top(\tilde{z}_{h-L},\tilde{a}_{h-L})] + \lambda_k I\right\} \\
&\qquad \cdot\left\{\int_{s_{h-L+1}\in\mathcal{S}}\omega(s_{h-L+1})\tilde{g}(s_{h-L+1})\,\mathrm{d}s_{h-L+1}\right\} \\
&\le k\mathbb{E}_{(\tilde{z}_{h-L},\tilde{a}_{h-L})\sim\beta_{h-L}}\left\{\left[\int_{s_{h-L+1}\in\mathcal{S}}\omega^\top(s_{h-L+1})\phi_{h-L+1}(\tilde{z}_{h-L},\tilde{a}_{h-L})\tilde{g}(s_{h-L+1})\,\mathrm{d}s_{h-L+1}\right]^2\right\} + B^2\lambda_k d \\
&= k\mathbb{E}_{(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})\sim\beta_{h-L}}\left\{\left[\int_{s_{h-L+1}\in\mathcal{S}}\omega^\top(s_{h-L+1})\phi_{h-L+1}(\tilde{z}_{h-L},\tilde{a}_{h-L})\tilde{g}(s_{h-L+1})\,\mathrm{d}s_{h-L+1}\right]^2\right\} + B^2\lambda_k d,
\end{aligned}
$$

where the inequality is because $\|g_h\|\le B$.

Moreover, we have

$$
\begin{aligned}
&k\mathbb{E}_{(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})\sim\beta_{h-L}}\left\{[\int_{s_{h-L+1}\in\mathcal{S}}\omega^\top(s_{h-L+1})\phi_{h-L+1}(\tilde{z}_{h-L},\tilde{a}_{h-L})\tilde{g}(s_{h-L+1})\,\mathrm{d}s_{h-L+1}]^2\right\} \\
&\le k\mathbb{E}_{(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})\sim\beta_{h-L}}\left\{[\mathbb{E}_{s_{h-L+1}\sim\mathbb{P}^{\mathcal{P}}_{h-L+1}(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})}\mathbb{E}^{\mathcal{P}}_{a_{h:h}\sim\mu^{\pi,h}}[g(z_h,a_h)\mid s_{h-L+1}]]^2\right\} + kB^2\epsilon_1 \\
&\le k\mathbb{E}_{(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})\sim\beta_{h-L},s_{h-L+1}\sim\mathbb{P}^{\mathcal{P}}_{h-L+1}(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L}),a_{h-L+1:h}\sim\mu^{\pi,h}}[g(z_h,a_h)\mid s_{h-L+1}]^2 + kB^2\epsilon_1 \\
&\le |A|^L k\mathbb{E}_{(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})\sim\beta_{h-L},a_{h-L+1:h}\sim U(\mathcal{A})}[g(z_h,a_h)]^2 + kB^2\epsilon_1 \\
&= |A|^L k\cdot\mathbb{E}_{(\tilde{z}_h,\tilde{a}_h)\sim\gamma_h}\{[g(\tilde{z}_h,\tilde{a}_h)]^2\} + kB^2\epsilon_1,
\end{aligned}
$$

where the first inequality is by Lemma 26, the second inequality is by Jensen's inequality and the last inequality is by importance sampling, the equation is by the definition of γ_h.

Then, the final statement is immediately concluded. □

Lemma 28 (L-step back inequality for the learned POMDP)**.** *Consider a set of functions $\{g_h\}_{h=0}^H$ that satisfies $g_h \in \mathcal{Z} \times \mathcal{A} \to \mathbb{R}$, s.t. $\|g_h\|_\infty \leq B$ for all $h \in [H]$. Then, for any policy π, we have*

$$\sum_{h=1}^{H} \mathbb{E}_{\pi}^{\widehat{\mathcal{P}}}[g(z_h, a_h)] \leq \sum_{h=1}^{H} \mathbb{E}_{z_{h-L-1}, a_{h-L-1} \sim \pi}^{\widehat{\mathcal{P}}}[\|\phi^\top(z_{h-L-1}, a_{h-L-1})\|_{\rho_{h-L-1}^{-1}, \phi_{h-L-1}}] \cdot \sqrt{|A|^L k \cdot \mathbb{E}_{(\tilde{z}_h, \tilde{a}_h) \sim \beta_h}^{\mathcal{P}}\{[g(\tilde{z}_h, \tilde{a}_h)]^2\} + B^2 \lambda_k d + k B^2 \epsilon_1} + B\epsilon_1.$$

Proof. For $h \in [H]$ and $h' \in [h-L+1, h]$, we define $\mathcal{X}_l = \mathcal{S}^l \times \mathcal{O}^l \times \mathcal{A}^{l-1}$ and

$$x_{h'} = (s_{h-L+1:h'}, o_{h-L+1:h'}, a_{h-L+1:h'-1}),$$

where $l = h' - h + L - 1$.

Now we define the moment matching policy $\mu^{\pi,h} = \{\mu_{h'}^{\pi,h} : \mathcal{X}_l \to \Delta(\mathcal{A})\}_{h'=h-L+1}^{h}$. We set $\mu^{\pi,h}$ as following:

$$\mu_{h'}^{\pi,h}(a_{h'} \mid x_{h'}) := \mathbb{E}_{\pi}^{\widehat{\mathcal{P}}}[\pi_{h'}(a_{h'} \mid z_{h'}) \mid x_{h'}] \text{ for } h' \leq h-1, \text{ and } \mu_h^{\pi,h} = \pi_h.$$

Then we define policy $\tilde{\pi}_h$ which takes first $h-L$ actions from π and remaining actions from $\mu^{\pi,h}$.

By Lemma B.2 in Efroni et al. (2022a), we have $d_h^{\widehat{\mathcal{P}},\pi}(z_h) = d_h^{\widehat{\mathcal{P}},\tilde{\pi}^h}(z_h)$. Then we have $d_h^{\widehat{\mathcal{P}},\pi}(z_h, a_h) = d_h^{\widehat{\mathcal{P}},\tilde{\pi}^h}(z_h, a_h)$ since $\mu_h^{\pi,h} = \pi_h$. Hence we have $\mathbb{E}_{\pi}^{\widehat{\mathcal{P}}}[g(z_h, a_h)] = \mathbb{E}_{\tilde{\pi}^h}^{\widehat{\mathcal{P}}}[g(z_h, a_h)]$.

Since $\mu^{\pi,h}$ is independent of s_{h-L+1}, we have the L-step-back decomposition:

$$\begin{aligned}
&\mathbb{E}_{\tilde{\pi}^h}^{\widehat{\mathcal{P}}}[g(z_h, a_h)] \\
&= \mathbb{E}_{o_{3-2L:h-L}, a_{3-2L,h-L} \sim \pi}^{\widehat{\mathcal{P}}}\left[\int_{s_{h-L+1}} \mathbb{P}(s_{h-L+1} \mid o_{3-2L:h-L}, a_{3-2L,h-L}) \cdot \mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}^{\widehat{\mathcal{P}}}[g(z_h, a_h) \mid s_{h-L+1}] \, \mathrm{d}s_{h-L+1}\right] \\
&\leq \mathbb{E}_{z_{h-L}, a_{h-L} \sim \pi}^{\widehat{\mathcal{P}}}\left[\int_{s_{h-L+1}} \phi^\top(z_{h-L}, a_{h-L}) \omega(s_{h-L+1}) \cdot \mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}^{\widehat{\mathcal{P}}}[g(z_h, a_h) \mid s_{h-L+1}] \, \mathrm{d}s_{h-L+1}\right] + B\epsilon_1 \\
&= \mathbb{E}_{z_{h-L}, a_{h-L} \sim \pi}^{\widehat{\mathcal{P}}} \phi^\top(z_{h-L}, a_{h-L}) \cdot \int_{s_{h-L+1} \in \mathcal{S}} \omega(s_{h-L+1}) \cdot \mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}^{\widehat{\mathcal{P}}}[g(z_h, a_h) \mid s_{h-L+1}] \, \mathrm{d}s_{h-L+1} + B\epsilon_1 \\
&\leq \mathbb{E}_{z_{h-L}, a_{h-L} \sim \pi}^{\widehat{\mathcal{P}}} \|\phi^\top(z_{h-L}, a_{h-L})\|_{\rho_{h-L}^{-1}, \phi_{h-L}} \\
&\qquad \cdot \left\| \int_{s_{h-L+1} \in \mathcal{S}} \omega(s_{h-L+1}) \cdot \mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}^{\widehat{\mathcal{P}}}[g(z_h, a_h) \mid s_{h-L+1}] \, \mathrm{d}s_{h-L+1} \right\|_{\rho_{h-L}, \phi_{h-L}} + B\epsilon_1,
\end{aligned}$$

where the first inequality is because Lemma 26.

Now we use $\tilde{g}(s_{h-L})$ to denote $\mathbb{E}_{a_{h-L+1:h} \sim \mu^{\pi,h}}^{\widehat{\mathcal{P}}}[g(z_h, a_h) \mid s_{h-L}]$ for notational simplification. We have

$$
\begin{aligned}
&\| \int_{s_{h-L+1}\in\mathcal{S}} \omega(s_{h-L+1})\tilde{g}(s_{h-L+1})\,\mathrm{d}s_h \|^2_{\rho_{h-L},\phi_{h-L}} \\
&= \left\{\int_{s_{h-L+1}\in\mathcal{S}} \omega(s_{h-L+1})\tilde{g}(s_{h-L+1})\,\mathrm{d}s_h\right\}^\top \\
&\qquad \cdot\left\{k\mathbb{E}_{(\tilde{z}_{h-L},\tilde{a}_{h-L})\sim\rho_{h-L+1}}[\phi_{h-L}(\tilde{z}_{h-L},\tilde{a}_{h-L})\phi^\top_{h-L}(\tilde{z}_{h-L},\tilde{a}_{h-L})]+\lambda_k I\right\} \\
&\qquad \cdot\left\{\int_{s_{h-L+1}\in\mathcal{S}} \omega(s_{h-L+1})\tilde{g}(s_{h-L+1})\,\mathrm{d}s_h\right\} \\
&\le k\mathbb{E}_{(\tilde{z}_{h-L},\tilde{a}_{h-L})\sim\rho_{h-L}}\left\{[\int_{s_{h-L+1}\in\mathcal{S}} \omega^\top(s_{h-L+1})\phi_{h-L+1}(\tilde{z}_{h-L},\tilde{a}_{h-L})\tilde{g}(s_{h-L+1})\,\mathrm{d}s_h]^2\right\}+B^2\lambda_k d \\
&= k\mathbb{E}_{(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})\sim\rho_{h-L}}\left\{[\int_{s_{h-L+1}\in\mathcal{S}} \omega^\top(s_{h-L+1})\phi_{h-L+1}(\tilde{z}_{h-L},\tilde{a}_{h-L})\tilde{g}(s_{h-L+1})\,\mathrm{d}s_h]^2\right\}+B^2\lambda_k d \\
&\le k\mathbb{E}_{(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})\sim\rho_{h-L}}\left\{[\mathbb{E}_{s_h\sim\widehat{\mathcal{P}}_h(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})}\mathbb{E}^{\widehat{\mathcal{P}}}_{a_{h:h}\sim\mu^{\pi,h}}[g(z_h,a_h)\mid s_{h-L+1}]]^2\right\}+B^2\lambda_k d+kB^2\epsilon_1,
\end{aligned}
$$

where the first inequality is because $\|g_h\|\le B$, the second inequality is by Lemma 26. Moreover, we have

$$
\begin{aligned}
&k\mathbb{E}_{(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})\sim\rho_{h-L}}\left\{[\mathbb{E}_{s_h\sim\widehat{\mathcal{P}}_h(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})}\mathbb{E}^{\widehat{\mathcal{P}}}_{a_{h:h}\sim\mu^{\pi,h}}[g(z_h,a_h)\mid s_{h-L+1}]]^2\right\} \\
&\le k\mathbb{E}_{(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})\sim\rho_{h-L}}\left\{[\mathbb{E}_{s_h\sim\mathbb{P}^{\mathcal{P}}_h(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})}\mathbb{E}_{a_{h:h}\sim\mu^{\pi,h}}[g(z_h,a_h)\mid s_{h-L+1}]]^2\right\} \\
&\le k\mathbb{E}_{(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})\sim\rho_{h-L},s_h\sim\mathbb{P}^{\mathcal{P}}_h(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L}),a_{h-L+1:h}\sim\mu^{\pi,h}}[g(z_h,a_h)\mid s_{h-L+1}]^2 \\
&\le |A|^L k\mathbb{E}_{(\tilde{o}_{3-2L:h-L},\tilde{a}_{3-2L:h-L})\sim\beta_{h-L},a_{h-L+1:h}\sim U(\mathcal{A})}[g(z_h,a_h)]^2 \\
&= |A|^L k\cdot\mathbb{E}_{(\tilde{z}_h,\tilde{a}_h)\sim\beta_h}\{[g(\tilde{z}_h,\tilde{a}_h)]^2\},
\end{aligned}
$$

where the first inequality is by Lemma 11, the second inequality is by Jensen's inequality, the last inequality is by importance sampling and the equation is by the definition of β_h.

Then, the final statement is immediately concluded. □

Lemma 29. *For any π, h, we have*

$$\|d^\pi_{\mathcal{P},h}-d^\pi_{\mathcal{M},h}\|_{TV}\le h\epsilon_1.$$

Proof. The proof is similar to Lemma D.1 in Agarwal et al. (2022). The only difference is that our condition is Proposition 8.

□

Next, we prove the almost optimism lemma restated below.

Lemma 30 (Almost Optimism for γ-observable POMDPs)**.** *Consider an episode k $(1\le k\le K)$ and set*

$$\alpha_k=\sqrt{k|\mathcal{A}|^L\zeta_k+4\lambda_k d+k\epsilon_1}/c,\quad \lambda_k=O(d\log(|\mathcal{F}|k/\delta)).$$

where c is an absolute constant. Conditioning on the event defined in Lemma 11, with probability $1-\delta$,

$$V^{\pi^*,\widehat{\mathcal{M}}_k,r+\widehat{b}^k}-V^{\pi^*,\mathcal{M},r}\ge-\frac{c\alpha_k L}{\sqrt{k}}-\mathcal{O}(H^2\epsilon_1)$$

holds for all $k\in[1,\cdots,K]$.

Proof. By Lemma 17, we have

$$\begin{aligned}
&V^{\pi^*,\widehat{\mathcal{M}},r+\widehat{b}} - V^{\pi^*,\mathcal{M},r}\\
&= \sum_{h=0}^{H-1} \mathbb{E}_{(z_h,a_h)\sim d^{\pi^*}_{\widehat{\mathcal{M}},h}}[\widehat{b}_h(z_h,a_h) + \mathbb{E}_{o'\sim\mathbb{P}_h^{\widehat{\mathcal{M}}}(\cdot|z_h,a_h)}[V_{h+1}^{\pi^*,\mathcal{M},r}(z'_{h+1})] - \mathbb{E}_{o'\sim\mathbb{P}_h^{\mathcal{M}}(\cdot|z_h,a_h)}[V_{h+1}^{\pi^*,\mathcal{M},r}(z'_{h+1})]]\\
&\geq \sum_{h=0}^{H-1} \mathbb{E}_{(z_h,a_h)\sim d^{\pi^*}_{\widehat{\mathcal{M}},h}}\Big[\min(c\alpha_k\|\widehat{\phi}_h(z,a)\|_{\Sigma^{-1}_{\rho_h,\widehat{\phi}_h}},1) + \mathbb{E}_{o'\sim\mathbb{P}_h^{\widehat{\mathcal{M}}}(\cdot|z_h,a_h)}[V_{h+1}^{\pi^*,\mathcal{M},r}(z'_{h+1})]\\
&\qquad - \mathbb{E}_{o'\sim\mathbb{P}_h^{\mathcal{M}}(\cdot|z_h,a_h)}[V_{h+1}^{\pi^*,\mathcal{M},r}(z'_{h+1})]\Big] \qquad (19)\\
&\geq \sum_{h=0}^{H-1} \mathbb{E}_{(z_h,a_h)\sim d^{\pi^*}_{\widehat{\mathcal{P}},h}}\Big[\min(c\alpha_k\|\widehat{\phi}_h(z,a)\|_{\Sigma^{-1}_{\rho_h,\widehat{\phi}_h}},1) + \mathbb{E}_{o'\sim\mathbb{P}_h^{\widehat{\mathcal{P}}}(\cdot|z_h,a_h)}[V_{h+1}^{\pi^*,\mathcal{M},r}(z'_{h+1})]\\
&\qquad - \mathbb{E}_{o'\sim\mathbb{P}_h^{\mathcal{P}}(\cdot|z_h,a_h)}[V_{h+1}^{\pi^*,\mathcal{M},r}(z'_{h+1})]\Big] - \mathcal{O}(H^2\epsilon_1), \qquad (20)
\end{aligned}$$

where in the fist inequality, we replace empirical covariance by population covariance by Lemma 16, here c is an absolute constant, the second inequality is by Lemma 29. We define

$$g_h(z,a) = \mathbb{E}_{o'_h\sim\mathbb{P}_h^{\widehat{\mathcal{P}}}(\cdot|z,a)}[V_{h+1}^{\pi^*,\mathcal{M},r}(c(z,a,o'_h))] - \mathbb{E}_{o'_h\sim\mathbb{P}_h^{\mathcal{P}}(\cdot|z,a)}[V_{h+1}^{\pi^*,\mathcal{M},r}(c(z,a,o'_h))]$$

Notice that we have $\|g_h\|_\infty \leq 1$. With Lemma 23, for any (z,a) we have

$$\mathbb{E}_{(z,a)\sim\rho_h}[g_h^2(z,a)] \leq \zeta_k,\ \mathbb{E}_{(z,a)\sim\beta_h}[g_h^2(z,a)] \leq \zeta_k \qquad (21)$$

By Lemma 28, we have:

$$\begin{aligned}
&\sum_{h=0}^{H-1} \mathbb{E}_{(z,a)\sim d^{\pi^*}_{\widehat{\mathcal{P}},h}}[g_h(z,a)] + \mathcal{O}(H^2\epsilon_1)\\
&\leq \sum_{h=1}^{H} \mathbb{E}^{\widehat{\mathcal{P}}}_{z_{h-L-1},a_{h-L-1}\sim\pi}\|\widehat{\phi}^\top(z_{h-L-1},a_{h-L-1})\|_{\rho^{-1}_{h-L-1},\widehat{\phi}_{h-L-1}}\\
&\qquad \cdot\sqrt{|A|^L k\cdot\mathbb{E}_{(\tilde{z}_h,\tilde{a}_h)\sim\beta_h}\{[g(\tilde{z}_h,\tilde{a}_h)]^2\} + B^2\lambda_k d + k\epsilon_1} + \mathcal{O}(H^2\epsilon_1),
\end{aligned}$$

where the first inequality is by Lemma 24 and the second inequality is by Lemma 21. Hence we have

$$\begin{aligned}
&\sum_{h=0}^{H-1} \mathbb{E}_{(z,a)\sim d^{\pi^*}_{\widehat{\mathcal{P}},h}}[g_h(z,a)]\\
&\leq \sum_{h=0}^{H-1} \min\{1, \mathbb{E}^{\widehat{\mathcal{P}}}_{z_{h-L-1},a_{h-L-1}\sim\pi}\|\widehat{\phi}^\top(z_{h-L-1},a_{h-L-1})\|_{\rho^{-1}_{h-L-1},\widehat{\phi}_{h-L-1}}\\
&\qquad \cdot\sqrt{|A|^L k\zeta_k + B^2\lambda_k d + k\epsilon_1}\} + \mathcal{O}(H^2\epsilon_1)\\
&\leq \sum_{h=0}^{H-1} \min\{1, c\alpha_k\mathbb{E}^{\widehat{\mathcal{M}}}_{z_{h-L-1},a_{h-L-1}\sim\pi}\|\widehat{\phi}^\top(z_{h-L-1},a_{h-L-1})\|_{\rho^{-1}_{h-L-1},\widehat{\phi}_{h-L-1}}\} + \mathcal{O}(H^2\epsilon_1), \qquad (22)
\end{aligned}$$

where the first inequality is by (21), in the last step we use Lemma 29 and the definition

$$\alpha_k = \sqrt{k|\mathcal{A}|^L\zeta_k + 4\lambda_k d + k\epsilon_1}/c.$$

For $h \leq 0$, we have

$$\|\widehat{\phi}_h^\top(z_h,a_h)\|_{\rho_h^{-1},\widehat{\phi}_h} = \sqrt{\frac{1}{k+\lambda}} < \frac{1}{\sqrt{k}}, \qquad (23)$$

since $\phi(s,a)=e_1$ for $h\le 0$.

Combine (8), (22) and (23), we conclude the proof. $\square$

Theorem 31. *With probability* $1-\delta$*, we have*

$$\sum_{k=1}^{K} V^{\pi^*,\mathcal{M},r}-V^{\pi^k,\mathcal{M},r}\le \mathcal{O}\big(H^2|\mathcal{A}|^L d^2 K^{1/2}\log(dK|\mathcal{F}|/\delta)^{1/2}+H^2 K d^{1/2}\log(dK|\mathcal{F}|/\delta)^{1/2}\epsilon_1\big).$$

Proof. The proof is similar to Theorem 19, we condition on the event that the MLE guarantee 23 holds, which happens with probability $1-\delta$. For fixed k we have

$$\begin{aligned}
&V^{\pi^*,\mathcal{M},r}-V^{\pi^k,\mathcal{M},r}\\
&\le V^{\pi^*,\mathcal{M},r+\widehat{b}^k}-V^{\pi^k,\mathcal{M},r}+\frac{c\alpha_k L}{\sqrt{k}}+\mathcal{O}(H^2\epsilon_1)\\
&\le V^{\pi^k,\widehat{\mathcal{M}},r+\widehat{b}^k}-V^{\pi^k,\mathcal{M},r}+\frac{c\alpha_k L}{\sqrt{k}}+\mathcal{O}(H^2\epsilon_1)\\
&=\sum_{h=0}^{H-1}\Big[\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{M}}}\big[\widehat{b}_h(z_h,a_h)+\mathbb{E}_{o_h'\sim\mathbb{P}_h^{\widehat{\mathcal{M}}}(o_h'|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{M}},r+\widehat{b}^k}(z_{h+1}')]\\
&\qquad-\mathbb{E}_{o_h'\sim\mathbb{P}_h^{\mathcal{M}}(o_h'|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{M}},r+\widehat{b}^k}(z_{h+1}')]\big]\Big]+\frac{c\alpha_k L}{\sqrt{k}}+\mathcal{O}(H^2\epsilon_1).
\end{aligned}$$

The first inequality comes from Lemma 30, the second inequality comes from $\pi^k=\operatorname{argmax}_\pi V^{\pi,\widehat{\mathcal{M}},r+\widehat{b}^k}$, and the last equation comes from Lemma 17.

By Lemma 29, we have

$$\begin{aligned}
&\sum_{h=0}^{H-1}\Big[\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{M}}}\big[\widehat{b}_h(z_h,a_h)+\mathbb{E}_{o_h'\sim\mathbb{P}_h^{\widehat{\mathcal{M}}}(o_h'|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{M}},r+\widehat{b}^k}(z_{h+1}')]\\
&\qquad-\mathbb{E}_{o_h'\sim\mathbb{P}_h^{\mathcal{M}}(o_h'|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{M}},r+\widehat{b}^k}(z_{h+1}')]\big]\Big]+\frac{c\alpha_k L}{\sqrt{k}}\\
&\le\sum_{h=0}^{H-1}\Big[\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}\big[\widehat{b}_h(z_h,a_h)+\mathbb{E}_{o_h'\sim\mathbb{P}_h^{\widehat{\mathcal{P}}}(o_h'|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{M}},r+\widehat{b}^k}(z_{h+1}')]\\
&\qquad-\mathbb{E}_{o_h'\sim\mathbb{P}_h^{\mathcal{P}}(o_h'|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{M}},r+\widehat{b}^k}(z_{h+1}')]\big]\Big]+\frac{c\alpha_k L}{\sqrt{k}}+\mathcal{O}(H^2\epsilon_1)
\end{aligned}$$

Denote

$$f_h(z_h,a_h)=\frac{1}{2H+1}\Big(\mathbb{E}_{o_h'\sim\mathbb{P}_h^{\widehat{\mathcal{P}}}(o_h'|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{M}},r+\widehat{b}^k}(z_{h+1}')]-\mathbb{E}_{o_h'\sim\mathbb{P}_h^{\mathcal{P}}(o_h'|z_h,a_h)}[V_{h+1}^{\pi^k,\widehat{\mathcal{M}},r+\widehat{b}^k}(z_{h+1}')]\Big).$$

Note that $\|\widehat{b}\|_\infty\le 1$, hence we have $\|V_{h+1}^{\pi^k,\widehat{\mathcal{M}},r+b}\|_\infty\le(2H+1)$. Combining this fact with the above expansion, we have

$$\begin{aligned}
V^{\pi^*,\mathcal{M},r}-V^{\pi^k,\widehat{\mathcal{M}},r}&=\sum_{h=0}^{H-1}\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}[\widehat{b}_h(z_h,a_h)]+(2H+1)\sum_{h=0}^{H-1}\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}[f_h(z_h,a_h)]\\
&\quad+\frac{c\alpha_k L}{\sqrt{k}}+\mathcal{O}(H^2\epsilon_1)\\
&\le\sum_{h=0}^{H-1}\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}[\widehat{b}_h(z_h,a_h)]+(2H+1)\sum_{h=0}^{H-1}\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}[f_h(z_h,a_h)]\\
&\quad+\frac{c\alpha_k L}{\sqrt{k}}+\mathcal{O}(H^2\epsilon_1).
\end{aligned}\tag{24}$$

First, we calculate the bonus term in (11). We have

$$\begin{aligned}&\sum_{h=0}^{H-1}\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}[\widehat{b}_h(z_h,a_h)]\\&\le\sum_{h=0}^{H-1}\mathbb{E}_{(\tilde{z},\tilde{a})\sim d_{h-L}^{\pi^k,\mathcal{P}}}\|\phi^*_{h-L}(\tilde{z},\tilde{a})\|_{\Sigma^{-1}_{\gamma_{h-L},\phi^*_{h-L}}}\sqrt{k|\mathcal{A}|^L\mathbb{E}_{(z,a)\sim\rho_h}[(\widehat{b}_h(z,a))^2]+4\lambda_k d+k\epsilon_1}+2H\epsilon_1,\end{aligned}$$

where the inequality is following Lemma 27 associate with $\|\widehat{b}_h\|_\infty\le 1$.

Note that we use the fact that $B=2$ when applying Lemma 27. In addition, we have that for any $h\in[H]$,

$$k\mathbb{E}_{(z,a)\sim\rho_h}[\|\widehat{\phi}_h(z,a)\|^2_{\Sigma^{-1}_{\rho_h,\widehat{\phi}_h}}]=k\,\mathrm{Tr}(\mathbb{E}_{\rho_h}[\widehat{\phi}_h\widehat{\phi}_h^\top]\{k\mathbb{E}_{\rho_h}[\widehat{\phi}_h\widehat{\phi}_h^\top]+\lambda_k I\}^{-1})\le d.$$

Then we have

$$\sum_{h=1}^{H}\mathbb{E}_{(z,a)\sim d_h^{\pi^k,\mathcal{P}}}[b_h(z,a)]\le\sum_{h=1}^{H}\mathbb{E}_{(\tilde{z},\tilde{a})\sim d_{h-L}^{\pi^k,\mathcal{P}}}\|\phi^*_{h-L}(\tilde{z},\tilde{a})\|_{\Sigma^{-1}_{\rho_{h-L},\phi^*_{h-L}}}\sqrt{|\mathcal{A}|^L\alpha_k^2 d+4\lambda_k d+k\epsilon_1}+2H\epsilon_1.$$

Now we bound the second term in (24). Following Lemma 27, with $\|f_h(z,a)\|_\infty\le 1$, we have

$$\begin{aligned}&\sum_{h=0}^{H-1}\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{P}}}[f_h(z_h,a_h)]\\&\le\sum_{h=0}^{H-1}\mathbb{E}_{(\tilde{z},\tilde{a})\sim d_{h-L}^{\pi^k,\mathcal{P}}}\|\phi^*_{h-L}(\tilde{z},\tilde{a})\|_{\Sigma^{-1}_{\gamma_{h-L},\phi^*_{h-L}}}\sqrt{k|\mathcal{A}|^L\mathbb{E}_{(z,a)\sim\rho_h}[f_h^2(z,a)]+4\lambda_k d+k\epsilon_1}\\&\le\sum_{h=0}^{H-1}\mathbb{E}_{(\tilde{z},\tilde{a})\sim d_{h-L}^{\pi^k,\mathcal{P}}}\|\phi^*_{h-L}(\tilde{z},\tilde{a})\|_{\Sigma^{-1}_{\gamma_{h-L},\phi^*_{h-L}}}\sqrt{k|\mathcal{A}|^L(\zeta_k)+4\lambda_k d+k\epsilon_1},\end{aligned}$$

where in the second inequality, we use $\mathbb{E}_{z,a\sim\rho_h}[f_h^2(z,a)]\le\zeta_k$. Then we have

$$\begin{aligned}&V^{\pi^*,\mathcal{M},r}-V^{\pi^k,\mathcal{M},r}\\&\le\sum_{h=0}^{H-1}\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{M}}}[b_h(z_h,a_h)]+(2H+1)\sum_{h=0}^{H-1}\mathbb{E}_{(z_h,a_h)\sim d_h^{\pi^k,\mathcal{M}}}[f_h(z_h,a_h)]+\frac{c\alpha_k L}{\sqrt{k}}+\mathcal{O}(H^2\epsilon_1)\\\le&\sum_{h=0}^{H-1}\mathbb{E}_{(\tilde{z},\tilde{a})\sim d_{h-L}^{\pi^k,\mathcal{P}}}\|\phi^*_h(\tilde{z},\tilde{a})\|_{\Sigma^{-1}_{\gamma_{h-L},\phi^*_{h-L}}}\sqrt{|\mathcal{A}|^L\alpha_k^2 d+4\lambda_k d+k\epsilon_1}\\&+(2H+1)\sum_{h=0}^{H-1}\mathbb{E}_{(\tilde{z},\tilde{a})\sim d_{h-L}^{\pi^k,\mathcal{P}}}\|\phi^*_{h-L}(\tilde{z},\tilde{a})\|_{\Sigma^{-1}_{\gamma_{h-L},\phi^*_{h-L}}}\sqrt{k|\mathcal{A}|^L\zeta_k+4\lambda_k d+k\epsilon_1}+\frac{c\alpha_k L}{\sqrt{k}}+\mathcal{O}(H^2\epsilon_1).\end{aligned}$$

Hereafter, we take the dominating term out. First, recall

$$\alpha_k=O\left(\sqrt{k|\mathcal{A}|^L\zeta_k+\lambda_k d}\right).$$

Second, recall that $\gamma_h^k(z,a) = \frac{1}{k}\sum_{i=0}^{k-1} d_h^{\pi^i}(z,a)$. Thus

$$\begin{aligned}
&\sum_{k=1}^K \mathbb{E}_{(\tilde z,\tilde a)\sim d_h^{\pi^k,\mathcal{P}}} \|\phi^*(\tilde z,\tilde a)\|_{\Sigma^{-1}_{\gamma_h^n,\phi_h^*}} \\
&\leq \sqrt{K\sum_{k=1}^K \mathbb{E}_{(\tilde z,\tilde a)\sim d_h^{\pi^k,\mathcal{P}}}[\phi_h^*(\tilde z,\tilde a)^\top \Sigma^{-1}_{\gamma_h^k,\phi_h^*}\phi_h^*(\tilde z,\tilde a)]} \\
&\leq \sqrt{K\bigg(\log\det\bigg(\sum_{k=1}^K \mathbb{E}_{(\tilde z,\tilde a)\sim d_h^{\pi^k,\mathcal{P}}}[\phi_h^*(\tilde z,\tilde a)\phi_h^*(\tilde z,\tilde a)^\top]\bigg) - \log\det(\lambda_1 I)\bigg)} \\
&\leq \sqrt{dK\log\bigg(1+\frac{K}{d\lambda_1}\bigg)},
\end{aligned}$$

where the first inequality is by Cauchy-Schwarz inequality, the second inequality is by Lemma 14 and the third inequality is by Lemma 15.

Finally, The MLE guarantee gives

$$\zeta_k = O\bigg(\frac{\log(dk|\mathcal{F}|/\delta)}{k}\bigg).$$

Combining all of the above, we have

$$\sum_{k=1}^K V^{\pi^*,\mathcal{M},r} - V^{\pi^k,\mathcal{M},r} \leq \mathcal{O}\big(H^2|\mathcal{A}|^L d^2 K^{1/2}\log(dK|\mathcal{F}|/\delta)^{1/2} + H^2Kd^{1/2}\log(dK|\mathcal{F}|/\delta)^{1/2}\epsilon_1)\big),$$

which concludes the proof. □

Lemma 32. *With probability* $1-\delta$, *we have*

$$V^{\pi^*,\mathcal{P},r} - V^{\pi^k,\mathcal{P},r} \leq \mathcal{O}\big(H^2|\mathcal{A}|^L d^2 K^{1/2}\log(dK|\mathcal{F}|/\delta)^{1/2} + H^2Kd^{1/2}\log(dK|\mathcal{F}|/\delta)^{1/2}\epsilon_1)\big).$$

Proof. Combined Lemma 9 and Theorem 31, we conclude the proof. □

E. Experiment Details

In our experiment, we assess the performance of the proposed algorithm on the partially-observed diabolical combination lock (pocomblock) problem, characterized by a horizon H and a set of 10 actions. At each temporal stage h, there exist three latent states $s_{i;h}$ for $i \in \{0,1,2\}$. We denote the states $s_{i;h}$ for $i \in \{0,1\}$ as favorable states and $s_{2;h}$ as unfavorable states. For each $s_{i;h}$ with $i \in \{0,1\}$, a specific action $a^*_{i;h}$ is randomly selected from the 10 available actions. When the agent is in state $s_{i;h}$ for $i \in \{0,1\}$ and performs action $a^*_{i;h}$, it transitions to states $s_{0;h+1}$ and $s_{1;h+1}$ with equal likelihood. Conversely, executing any other actions will deterministically lead the agent to state $s_{2;h+1}$. In the unfavorable state $s_{2;h}$, any action taken by the agent will inevitably result in transitioning to state $s_{2;h+1}$.

As for the reward function, a reward of 1 is assigned to states $s_{i;H}$ for $i \in \{0,1\}$, signifying that favorable states at horizon H yield a reward of 1. Additionally, with a 0.5 probability, the agent receives an anti-shaped reward of 0.1 upon transitioning from a favorable state to an unfavorable state. All other states and transitions yield a reward of zero. When the step h is odd, the observation o is generated with a dimension of $2^{\lceil\log(H+4)\rceil}$ by concatenating the one-hot vectors of latent state s and horizon h, introducing Gaussian noise sampled from $\mathcal{N}(0,0.1)$ for each dimension, appending 0 if needed, and multiplying with a Hadamard matrix. For even steps h, the observations corresponding to one of the good states and the bad states are identical, the other good state's observation function is the same as the time step is odd. The initial state distribution is uniformly distributed across $si;0$ for $i \in \{0,1\}$. We employ a two-layer neural network to capture the essential features of the problem.

It is noteworthy that the optimal policy consists of selecting the specific action $a^*_{i;h}$ at each step h. Once the agent enters an unfavorable state, it remains trapped in the unfavorable state for the entire episode, thus failing to attain the substantial reward signal at the conclusion. This presents an exceptionally demanding exploration problem, as a uniform random policy offers a mere 10^{-H} probability of achieving the objectives. In our experiment, we compare our method with BRIEE, the latest representation learning algorithm for MDP. In particular, we make modifications to BRIEE to take a sequence of observations as input for representation learning, in order to work in the POMDP settings.

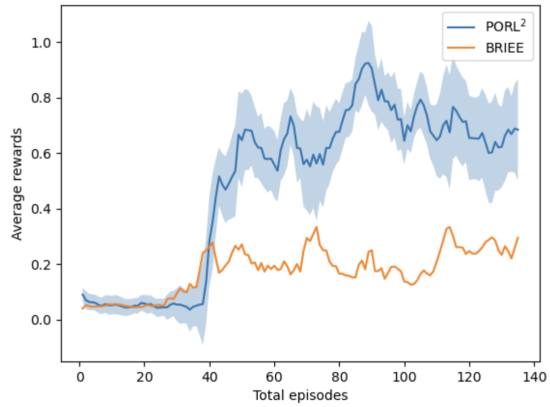

Figure 2. Moving average of evaluation returns of pocomblock for PORL^2 and BRIEE

Implementation details for PORL^2 In our implementation, we extend the BRIEE framework by considering two consecutive observations (pocomblock is 2-decodable), o_{h-1} and o_h, as a single variable, denoted as z_h. When employing LSVI-LLR, we compute the value function based on the state instead of z. This choice is justified by the fact that the rewards associated with different observations of the same state are equal. Figure 2 is the moving average of evaluation returns of pocomblock for PORL^2 and BRIEE.

We record the hyperparameters we try and the final hyperparameter we use for PORL^2 in Table 2 and BRIEE in Table 3.

	Value Considered	Value
Batch size	{256, 512}	512
Discriminator f number of gradient steps	{4,8,16,32}	8
Horizon	{4,6,7,10}	7
Seeds	{1,12,123,1234,12345}	12345
Decoder ϕ number of gradient steps	{4,8,16,32}	8
The number of iterations of representation learning	{6,8,10,12}	10
LSVI-LLR bonus coefficient β	{1}	1
LSVI-LLR regularization coefficient λ	{1}	1
Optimizer	{ SGD }	SGD
Decoder ϕ learning rate	{1e-3,1e-2}	1e-2
Discriminator f learning rate	{1e-3,1e-2}	1e-2

Table 2. Hyperparameters for PORL^2

	Value Considered	Value
Batch size	$\{256, 512\}$	512
Discriminator f number of gradient steps	$\{4,8,16,32\}$	8
Horizon	$\{4,6,7,10\}$	7
Seeds	$\{1,12,123,1234,12345\}$	12345
Decoder ϕ number of gradient steps	$\{8\}$	8
The number of iterations of representation learning	$\{10\}$	10
LSVI-LLR bonus coefficient β	$\{1,10\}$	10
LSVI-LLR regularization coefficient λ	$\{1\}$	1
Optimizer	{ SGD }	SGD
Decoder ϕ learning rate	$\{$1e-2$\}$	1e-2
Discriminator f learning rate	$\{$1e-2$\}$	1e-2

Table 3. Hyperparameters for BRIEE