Provably Efficient Offline Reinforcement Learning for Partially Observable Markov Decision Processes

Hongyi Guo [1] Qi Cai [1] Yufeng Zhang [1] Zhuoran Yang [2] Zhaoran Wang [1]

Abstract

We study offline reinforcement learning (RL) for partially observable Markov decision processes (POMDPs) with possibly infinite state and observation spaces. Under the undercompleteness assumption, the transition kernel can be estimated via solving a series of confounded regression problems. To solve the confounding problem, we select a proper instrumental variable (IV) and solves the IV regression problem to construct confidence regions for the model parameters. We get the final policy via pessimistic planning within the confidence regions. We prove that the proposed algorithm attains an ϵ-optimal policy using an offline dataset containing $\widetilde{\mathcal{O}}(1/\epsilon^2)$ episodes, provided that the behavior policy has good coverage over the optimal trajectory. To our best knowledge, our algorithm is the first provably sample efficient offline algorithm for POMDPs that is not tabular.

1. Introduction

In the past few years, deep reinforcement learning (RL) has shown its great potential to achieve human-level intelligence (Mnih et al., 2015; Silver et al., 2016; 2017; Vinyals et al., 2019; Ye et al., 2020; Wei et al., 2018; Pathak et al., 2019). In most literature, the environment is modeled as a Markov decision process (MDP) (Sutton & Barto, 2018), where the agent can access a state that contains all the information of the whole system for action selection. A bunch of algorithms are proposed and proved theoretically to have strong performance guarantees on MDPs, in terms of regret or sample complexity (Fan et al., 2020; Cai et al., 2019; Liu et al., 2019a; Wang et al., 2019; Jin et al., 2020b; Cai et al., 2020). However, the assumption of full observability of the state is often violated in practice, which restricts the application of RL to scenarios with well-defined MDPs like video games (Mnih et al., 2015; Vinyals et al., 2019; Ye et al., 2020). In most cases, the agent can only observe a partial state or a noisy version of the true state. To model such situations, the partially observable Markov decision process (POMDP) (Sondik, 1971; Spaan, 2012) is introduced.

In order to compensate the information loss in the partial observations, action selections in POMDPs should depend on the whole history instead of the observation it alone. Such a dependency leads to huge computing burden since processing a series of observations and actions in a POMDP needs more careful design and consumes much more time than processing one state in an MDP, especially when the episode is long. Also, since the model parameters of a POMDP can not be estimated directly from the data due to the latency of the states, we can either infer the latent state or estimate the model parameters from history. Both leads to great challenges compared with learning in an MDP.

Despite those challenges, many algorithms have been proposed to tackle the long dependency and are proved to have a sound sample complexity or regret bound Azizzadenesheli et al. (2016); Guo et al. (2016); Kwon et al. (2021); Jin et al. (2020a); Xiong et al. (2021). Those algorithms focus on online learning settings, switching between sample collection and strategy improvement during training. Online algorithms can acquire new data from interacting with the environment, and the distribution of those data can be altered towards a desirable direction by modulating the strategy of interaction. Thus, online algorithms are often sample efficient. However, in scenarios like healthcare (Yu et al., 2021; Tang & Wiens, 2021; Sonabend-W et al., 2020) and autonomous driving (Kiran et al., 2021; Shi et al., 2021b), it's impractical to collect new data because of the inconvenience or even potential danger caused by a bad interaction strategy. Thus, in those settings, offline learning is a more desirable way. Though sample efficient RL algorithms have been proposed for offline setting (Jin et al., 2021; Zanette et al., 2021; Kidambi et al., 2020; Yu et al., 2020), there is no theoretical guarantees of any offline RL algorithm on a POMDP. In this work, we combine offline reinforcement

[*] Equal contribution [1] Department of Industrial Engineering and Management Sciences, Northwestern University [2] Department of Statistics and Data Science, Yale University. Correspondence to: Zhaoran Wang <zhaoranwang@gmail.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

learning with POMDP framework and aim to answer the following question,

Can offline reinforcement learning be provably efficient on POMDPs?

To answer this question, we propose a new pessimistic offline RL algorithm. Specifically, we consider a broad class of linear POMDPs where both the observation emission kernel and the state transition kernel are linear in the feature mappings, which allows the state and observation space to be arbitrarily large or continuous. A similar model has been studied in (Yang & Wang, 2020) for MDPs. Our algorithm exploits the undercompleteness assumption and the pessimism principle in the face of uncertainty. We formulate the learning of the model parameters as a confounded regression problem, and select proper instrumental variables to cancel the correlation between the covariate and the perturbation. We theoretically prove that, for any $\epsilon > 0$, our algorithm obtains an ϵ-optimal policy with a proper offline dataset containing $\mathrm{poly}(H, |\mathcal{A}|, d_\phi, d_\psi) \cdot \widetilde{\mathcal{O}}(1/\epsilon^2)$ episodes, where H is the episode length, $|\mathcal{A}|$ is the number of available actions, d_ϕ and d_ψ are dimensions of the feature mappings for observations and states, respectively. To our best knowledge, our algorithm is the first provably sample efficient offline RL algorithm for POMDPs that is not tabular.

The remainder of this paper is organized as follows. We introduce related literature in §1.1 and notations in §1.2. The preliminaries are given in §2. We propose our algorithm in §3. In §4, we introduce our main theoretical results. We sketch the proof of our main theorem in §5.

1.1. Related Work

POMDPs. Our work is related to the literature on reinforcement learning on POMDPs. See (Spaan, 2012) and the references therein. Recently, a polynomial sample complexity for learning of the model parameters is achieved in Azizzadenesheli et al. (2016); Xiong et al. (2021); Guo et al. (2016), which all builds on the estimation of the parameters of hidden Markov models (HMMs) using spectral methods (Anandkumar et al., 2012; 2014). From the perspective of POMDPs, an HMM is a special case with a fixed action sequence. To cope with this difference, Azizzadenesheli et al. (2016) restrict their oracle to the memoryless policy that only depends on the current observation instead of using all historical information to form the belief of the underlying state. Closest to our work is Jin et al. (2020a). They also propose to make an undercompleteness assumption, where they assume the latent states are less than the possible observations. We extend this assumption to a broader class of POMDPs instead of tabular cases in Jin et al. (2020a) and introduce a more sophisticated undercompleteness assumption. Instead of estimating the real POMDP model, (Jin et al., 2020a) utilizes the observable operator model (Jaeger, 2000) induced by the environment, which is the same idea as ours.

Causal Inference. There has been a line of works that apply methods in Causal Inference literature to identify the POMDPs. The unobserved states admits unmeasured confounders in the POMDP model. Shi et al. (2021a); Bennett & Kallus (2021) study the Off-Policy Evaluation (OPE) problems in POMDPs and identify the POMDP with bridge functions. The goal of OPE in POMDPs is to estimate the value of an evaluation policy, which is a function of observed variables, using the data generated by a behavior policy. The most related work is Liao et al. (2021) that identifies the confounded transition dynamics via instrumental variables and proposes IV-aided Value Iteration (IVVI) to recover the optimal policy.

Offline RL. In the context of offline RL, our work is related to the recent pessimistic or conservative RL algorithms (Zanette et al., 2021; Jin et al., 2021; Kumar et al., 2020; Kidambi et al., 2020; Yu et al., 2020; Buckman et al., 2020) which tries to find the best policy in the face of uncertainty given an offline dataset. The intuition behind pessimism is the need of worst-case guarantees arisen from the concern about insufficient exploration in the offline dataset. Specifically, Jin et al. (2021); Yu et al. (2020) adds bonus to the rewards which penalize states not well-covered by the dataset. The works of Zanette et al. (2021); Kidambi et al. (2020) construct an MDP model on which the performance of any policy lower-bounds that on the real environment, and then learn a near-optimal policy on this model. Our idea is closest to the second approach.

Function Approximations. Our linear POMDP assumption draws connection to the rich literature on RL with function approximations (Yang & Wang, 2020; Jin et al., 2020b; 2021; Cai et al., 2020; Wang et al., 2020a; Zhou et al., 2021; Ayoub et al., 2020). Among those works, our model is most related to the linear factored MDPs (Yang & Wang, 2020) with known feature mappings and an unknown kernel matrix. However, the approach in Yang & Wang (2020) can not be applied to POMDP directly because of the latency of the states.

1.2. Notation

For any $n \in \mathbb{N}$, we denote $[n] = \{1, 2, \ldots, n\}$. We denote by $\|\cdot\|_p$ the ℓ^p-norm of a vector or L^p norm of a function. For any operator M, we denote by $\|M\|_{p\mapsto q}$ the operator norm of M induced by the ℓ^p-norm or L^p-norm of the domain and ℓ^q or L^q-norm of the range. For any discrete or continuous set $\mathcal{X}$, we denote by $L^p(\mathcal{X})$ the L^p space of functions over $\mathcal{X}$ and $\Delta(\mathcal{X})$ the space of probability density functions over $\mathcal{X}$ when $\mathcal{X}$ is continuous or the space of probability mass functions when $\mathcal{X}$ is discrete. Also, the notation of the integral $\int_{\mathcal{X}}$ in this paper is general for summation over $\mathcal{X}$ no matter $\mathcal{X}$ is discrete or continuous. For a sequence of variables $x_1, x_2, \ldots$, we use $x_{i:j}$ to denote the subsequence $x_i, x_{i+1}, \ldots, x_j$ for $i \leq j$. Lastly, we use the notation $\mathrm{linspan}(\cdot)$ to represent the linear span.

2. Preliminary

We consider episodic Partially Observable Markov Decision Process $\mathcal{M}(H, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{E}, r, \mu_1)$. Here, H is the horizon, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{O}$ is the observation space, and μ_1 is the initial state distribution. We consider finite and discrete action space but allow the state space and observation space to be arbitrarily large or continuous. The transition kernel $\mathcal{T} = \{\mathcal{T}_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})\}_{h=1}^H$ defines the transition probabilities at state $s \in \mathcal{S}$ and when action $a \in \mathcal{A}$ is taken. The emission $\mathcal{E} = \{\mathcal{E}_h : \mathcal{S} \to \Delta(\mathcal{O})\}_{h=1}^H$ specifies the distribution over observations for any state $s \in \mathcal{S}$ at any step $h \in [H]$. Specifically, we assume the reward functions $r = \{r_h : \mathcal{O} \times \mathcal{A} \to [0,1]\}_{h=1}^H$ are given and the rewards are normalized.

In POMDPs, the agent has no access to the state of the system. Instead, only the observation for that state is revealed to the agent. At the beginning of each episode, an initial state s_1 is sampled from μ_1. At each step $h \in [H]$, the agent observes $o_h \in \mathcal{O}$, which is sampled from $\mathcal{E}_h(\cdot \,|\, s_h)$, where s_h is the latent state. Then, the agent picks an action $a_h \in \mathcal{A}$ by following its policy. Also, the agent receives reward $r_h(o_h, a_h)$ for reaching the current observation o_h and taking action a_h. Afterwards, the system transits to the next latent state $s_{h+1} \sim \mathcal{T}(\cdot \,|\, s_h, a_h)$ following the transition kernel. The process continues until reaching the terminating state s_{H+1}.

To deal with the partial observability, the agent takes into account all the history observations and actions when taking an action. That is to say, $\pi = \{\pi_h : \Gamma_h \to \Delta(\mathcal{A})\}_{h=1}^H$, where $\Gamma_h = (\mathcal{O} \times \mathcal{A})^{h-1} \times \mathcal{O}$ is the space of all the histories at step h. We denote by Π the space of all policies, which is known in prior. To describe the environment, we parameterize the POMDP and use $\theta \in \Theta$ to denote all the unknown parameters in it, where Θ is the parameter space given in prior, which will be specified later in Assumption 2.1. Then, we use $\mathcal{T}^\theta$ and $\mathcal{E}^\theta$ to denote the parameterized versions of $\mathcal{T}$ and $\mathcal{E}$.

Similar to the MDP case, here we also aim to find the policy $\pi^* \in \Pi$ that maximizes the expected accumulated reward. Since the horizon is finite and the actions are discrete and finite, there always exists an optimal policy π^* such that

$$\pi^* = \operatorname*{argmax}_{\pi \in \Pi} \mathcal{J}(\theta^*, \pi), \tag{2.1}$$

$$\text{where} \quad \mathcal{J}(\theta, \pi) = \mathbb{E}_{\theta,\pi}\Big[\sum_{h \in [H]} r_h(o_h, a_h)\Big] \tag{2.2}$$

for any $\theta \in \Theta$ and $\pi \in \Pi$. Here, the symbols θ and π on the subscripts of the expectations indicate that the parameter of the underlying POMDP is θ, and all the actions are selected by the policy π.

We consider the offline reinforcement learning setting with a dataset $\mathcal{D}_N = \{(o_h^n, a_h^n, r_h^n, \rho_h^n)\}_{(n,h) \in [N] \times [H]}$ that contains N trajectories collected by some behavior policy $\bar{\pi}$. Specifically, besides the action a_h^n in each step, we also require the dataset to contain the probability of the behavior policy to take that action, denoted by ρ_h^n. Our goal is to learn an ϵ-optimal policy π with $\mathcal{D}_N$. The suboptimality of any policy π is characterized by the following regret

$$\mathrm{Regret}(\pi) = \mathcal{J}(\theta^*, \pi^*) - \mathcal{J}(\theta^*, \pi), \tag{2.3}$$

which is defined as the gap between the expected accumulated reward by following the optimal policy π^* and that by following π in the actual POMDP.

2.1. Linear POMDP

We extend the linear MDP assumptions (Jin et al., 2020b; 2021; Wang et al., 2020a; Zhou et al., 2021; Ayoub et al., 2020; Yang & Wang, 2020) to POMDPs and consider the POMDPs where both the transition kernel and the emission kernel can be fully embedded in the given feature spaces, as stated in the following assumption.

Assumption 2.1. [Linear POMDP] For any $h \in [H]$, there exists known feature mappings

$$\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^{d_\phi},\ \psi : \mathcal{S} \to \mathbb{R}^{d_\psi},\ \varphi : \mathcal{O} \to \mathbb{R}^{d_\varphi},$$

where the entries of ψ and φ are probability distributions. Let $\theta = (\{M_h^\theta\}_{h \in [H]}, \{\eta_h^\theta\}_{h \in [H]})$ be all the unknown parameters in the POMDP, including matrices $M_h^\theta \in \mathbb{R}^{d_\phi \times d_\psi}$ and feature mappings $\eta_h^\theta : \mathcal{S} \to \mathbb{R}^{d_\eta}$. For a given parameter space $\Theta \subset \mathbb{R}^{d_\phi \times d_\psi} \times L^{d_\eta}(\mathcal{S})$, there exists $\theta^* \in \Theta$ such that

$$\mathcal{T}_h(s' \,|\, s, a) = \phi(s,a)^\top M_h^{\theta^*} \psi(s'),$$
$$\mathcal{E}_h(\cdot \,|\, s) = \varphi(o)^\top \eta_h^{\theta^*}(s)$$

for any $h \in [H]$, $s, s' \in \mathcal{S}$ and $a \in \mathcal{A}$.

The assumption of the linear POMDP requires that both emission function and transition function can be linearly approximated by known basis ψ and φ. For notational simplicity, we denote by ψ_i and φ_j the i-th entry of ψ

and the j-th entry of φ for any $i \in [d_\psi]$ and $j \in [d_\varphi]$, respectively. The same transition kernel has been defined in Yang & Wang (2020) for MDPs. However, their algorithm can not be applied to POMDPs since the latency of the states. For any $\theta = (\{M_h^\theta\}_{h\in[H]}, \{\eta_h^\theta\}_{h\in[H]})$, we define the following state transition function and observation emission function

$$\mathcal{T}_h^\theta(s' \,|\, s, a) = \phi(s,a)^\top M_h^\theta \psi(s'), \tag{2.4}$$
$$\mathcal{E}_h^\theta(\cdot \,|\, s) = \varphi(o)^\top \eta_h^\theta(s). \tag{2.5}$$

The linearity of the transition kernel and the emission kernel helps us to characterize undercompleteness (Jin et al., 2020a). In tabular settings, it means the states are less than the observations. Prior to introduce undercompleteness to our linear POMDP, we define the observation operator $\mathbb{O}_h^\theta : L^1(\mathcal{S}) \to L^1(\mathcal{O})$ for any $(\theta, h) \in \Theta \times [H]$ and function $f \in L^1(\mathcal{S})$ by

$$(\mathbb{O}_h^\theta f)(o) = \int_{\mathcal{S}} \mathcal{E}_h^\theta(o \,|\, s) f(s) \,\mathrm{d}s, \tag{2.6}$$

where $o \in \mathcal{O}$ is arbitrarily. When f is a distribution of the state s_h, such an operator maps it to the distribution of its observation o_h under the emission kernel $\mathcal{E}_h^\theta$. In tabular settings, such an operator is defined by matrix multiplication. Thus, the inverse operator is characterized by the Moore-Penrose inverse of the matrix multiplication (Jin et al., 2020a). However, in our linear POMDP with function approximations, the definition of the inverse operator becomes more sophisticated, which is described in the following assumption.

Assumption 2.2. [Undercompleteness] For any $h \in [H]$ and $\theta \in \Theta$, there exists function $\xi_h^\theta : \mathcal{S} \times \mathcal{O} \to \mathbb{R}$ and corresponding operator $\mathbb{U}_h^\theta : L^1(\mathcal{O}) \to L^1(\mathcal{S})$ such that

$$\mathbb{U}_h^\theta \mathbb{O}_h^\theta f = f, \quad \|\mathbb{U}_h^\theta\|_{1\mapsto 1} \le \gamma \tag{2.7}$$

for any $f \in \text{linspan}(\psi)$, where $\gamma > 0$ is an absolute constant and the operator $\mathbb{U}_h^\theta$ is defined by

$$(\mathbb{U}_h^\theta g)(s) = \int_{\mathcal{O}} \xi_h^\theta(s, o) g(o) \,\mathrm{d}o \tag{2.8}$$

for any $g \in L^1(\mathcal{O})$.

Assumption is an extension of the undercompleteness assumption in tabular POMDPs defined in Jin et al. (2020a). It defines the left inverse of the observation operator $\mathbb{O}$. Recall that $\mathbb{O}$ maps a distribution of the states to that of the corresponding observations. Then, by (2.6), we can infer the distribution of the states from that of the observations. Though we can not directly acquire the latent state from the observation, the distribution of the state still carries effective information To compare with Jin et al. (2020a) which requires the number of states is smaller than that of the observations, here we only require the linear mappings $\mathbb{O}$ are injective. Interested readers can find more discussion in §A.3. In what follows, we build our algorithm and analysis on this assumption.

3. Algorithm

3.1. General Framework

To develop an algorithm for POMDPs, our strategy works as follows. Given the offline dataset, we first estimate the model parameter θ^*. Instead of having an exact estimator, we construct a confidence region $\widehat{\Theta} \subset \Theta$, which should contain θ^* with high probability. With such a confidence region, for any policy $\pi \in \Pi$, we obtain a value function estimator which lower-bounds the true value function of π. Then, it's natural to maximize those lower bounds over some family Π of policies, which leads to the saddle-point problem

$$(\widehat{\pi}, \widehat{\theta}) = \operatorname*{argmax}_{\pi\in\Pi} \operatorname*{argmin}_{\theta\in\widehat{\Theta}} \mathcal{J}(\theta, \pi), \tag{3.1}$$

where $\mathcal{J}$ is the expected total reward defined in (2.2). Note that the method to solve (3.1) is out of the scope of our study.talk about it later. The problem of our interest is how to estimate the model parameter θ^* and construct the confidence region $\widehat{\Theta}$ with the offline dataset collected by some behavior policy $\overline{\pi}$.

The most important part of identifying the model parameters is to identify the transition kernel. We assume all other parameters, including the reward function and the initial state distribution, are given in prior. In MDPs, to identify the transition kernel, it suffices to collect (s_h, a_h, s_{h+1}) pairs and the transition function can be identified by $\mathcal{T}_h^\theta(s' \,|\, s, a) = \mathbb{E}[\mathbb{1}\{S_{h+1} = s'\} \,|\, S_h = s, A_h = a]$. Thus, we can estimate the transition kernel sufficiently accurate as long as the dataset is large enough and provides a uniform coverage over all the state-action pairs.

However, in POMDPs, the states are unobserved. We need to find other way to identify the transition kernel. We start by defining the following two random functions

$$X_{h,a}(o) = \mathbb{1}^{\overline{\pi}}_{\mathrm{do}(A_{h-1}=a)}\{O_h = o\},$$
$$Y_{h,a,a'}(o, o') = \mathbb{1}^{\overline{\pi}}_{\mathrm{do}(A_{h-1:h}=(a,a'))}\{O_{h:h+1} = (o, o')\},$$

for any $(h, a, a', o, o') \in [H] \times \mathcal{A}^2 \times \mathcal{O}^2$. Here, the superscript $\overline{\pi}$ implies that we are following the behavior policy $\overline{\pi}$, whereas the do-operation in the subscript performs a deterministic intervention on the policy $\overline{\pi}$ at specified steps while leaving all other actions untouched. For instance, in the definition of $Y_{h,a,a'}$, we assign a and a' to A_{h-1} and A_h, respectively, regardless of the observations, while the actions $A_{1:h-2}$ are still taken by the behavior policy $\overline{\pi}$ based on the observations. Such a do-operation is commonly adopted in Causal Inference (Pearl, 2009).

Intuitively, if there exists linear operators $\widetilde{\mathbb{F}}^\theta_{h,a} : (\mathcal{O} \to \mathbb{R}) \to (\mathcal{O}^2 \to \mathbb{R})$ associated with the model parameter θ such that

$$Y_{h,a,a'} = \widetilde{\mathbb{F}}^{\theta^*}_{h,a'} X_{h,a} + U_{h,a,a'},$$

where $U_{h,a,a'}$ is the perturbation term independent of $X_{h,a}$ and has an expectation of 0, then we can regress $Y_{h,a,a'}$ on $X_{h,a}$ for any $(h,a,a') \in [H] \times \mathcal{A}^2$ and estimate the model parameter θ via estimating the operator $\widetilde{\mathbb{F}}^\theta_{h,a'}$ with observations in the dataset. To that end, we define $\widetilde{\mathbb{F}}^\theta_{h,a}$ to be the linear operator

$$(\widetilde{\mathbb{F}}^\theta_{h,a} f)(o,o') = \int_{\mathcal{O}} f(\widetilde{o}) \cdot \mathcal{F}^\theta_{h,a}(\widetilde{o}, o, o')\,\mathrm{d}\widetilde{o}, \tag{3.2}$$

where the integral kernel $\mathcal{F}^\theta_{h,a}$ is given by

$$\begin{aligned}\mathcal{F}^\theta_{h,a}(o,o',o'') = \int_{\mathcal{S}} &\xi^\theta_h(s,o) \cdot \\ &\mathbb{P}^\theta(O_{h:h+1} = (o',o'') \,|\, S_h = s, A_h = a)\,\mathrm{d}s,\end{aligned} \tag{3.3}$$

where ξ^θ_h is defined in Assumption 2.2. The operator $\widetilde{\mathbb{F}}^\theta_{h,a}$ is only influenced by the model parameter θ and is independent of any policy. With such a linear operator, we have the following lemma which characterizes the perturbation term.

Lemma 3.1. *For any $(h,a,a') \in [H] \times \mathcal{A}^2$, we define random function*

$$\widetilde{U}_{h,a,a'} = Y_{h,a,a'} - \mathbb{E}[Y_{h,a,a'} \,|\, O_{1:h-1}].$$

Then, there exists a linear operator $\mathbb{L}^\theta_{h,a} : (\mathcal{O}^2 \to \mathbb{R}) \to (\mathcal{O}^2 \to \mathbb{R})$ such that

$$Y_{h,a,a'} = \mathbb{F}^{\theta^*}_{h,a'} X_{h,a} + \mathbb{L}^{\theta^*}_{h,a'} \widetilde{U}_{h,a,a'}. \tag{3.4}$$

Proof. See §A.1 for a detailed proof. □

Lemma 3.1 gives the closed form of the perturbation term $U_{h,a,a'} = \mathbb{L}^\theta_{h,a'} \widetilde{U}_{h,a,a'}$. By this definition, we have the desirable result $\mathbb{E}[U_{h,a,a'}(o,o',o'')] = 0$ for any $(h,a,a',o,o') \in [H] \times \mathcal{A}^2 \times \mathcal{O}^2$. However, this perturbation term $U_{h,a,a'}$ influences both $X_{h,a}$ and $Y_{h,a,a'}$. In the language of Causal Inference (Pearl, 2009), we call $U_{h,a,a'}$ a confounder of $X_{h,a}$ and $Y_{h,a,a'}$. The existence of such a confounder may lead us to biased estimations of $\widetilde{\mathbb{F}}^\theta_{h,a'}$ as well as θ. To fix this issue, we draw inspirations from the method of instrumental variables (IV). An instrumental variable must be correlated with the covariate $X_{h,a}$ but uncorrelated with the perturbation $U_{h,a,a'}$. We illustrate the relationship between $X_{h,a}$, $Y_{h,a,a'}$, $U_{h,a,a'}$ and Z_h in Figure 1. In our case, we introduce the following instrumental variable

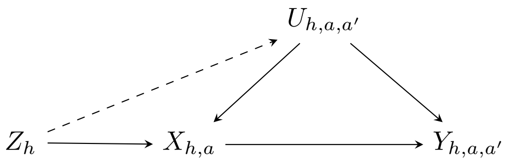

Figure 1. The relationship between $X_{h,a}$, $Y_{h,a,a'}$, $U_{h,a,a'}$, and Z_h. The arrows indicates the dependency between those variables. The dashed arrow indicates two uncorrelated variables. In this figure, $U_{h,a,a'}$ affects both $X_{h,a}$ and $Y_{h,a,a'}$ directly, Z_h only affects $X_{h,a}$ directly, and Z_h and $U_{h,a,a'}$ are uncorrelated.

$$Z_h = O_{h-1}. \tag{3.5}$$

To verify that Z_h is uncorrelated with $U_{h,a,a'}$, we note that

$$\begin{aligned}&\mathbb{E}[U_{h,a,a'} \,|\, Z_h] = \mathbb{L}^\theta_{h,a'} \mathbb{E}[\widetilde{U}_{h,a,a'} \,|\, Z_h] \\ &= \mathbb{L}^\theta_{h,a'}(\mathbb{E}[Y_{h,a,a'} \,|\, O_{h-1}] - \mathbb{E}[\mathbb{E}[Y_{h,a,a'} \,|\, O_{1:h-1}] \,|\, O_{h-1}]) \\ &= \mathbb{L}^\theta_{h,a'}(\mathbb{E}[Y_{h,a,a'} \,|\, O_{h-1}] - \mathbb{E}[Y_{h,a,a'} \,|\, O_{h-1}]) = 0,\end{aligned}$$

where the first equality holds because the operator $\mathbb{L}^\theta_{h,a'}$ is linear, and the last equality follows from the tower property of expectations. With the help of the instrumental variable Z_h, it holds for any (h,a,a') that

$$\begin{aligned}\mathbb{E}[Y_{h,a,a'} \,|\, Z_h] &= \widetilde{\mathbb{F}}^\theta_{h,a'} \mathbb{E}[X_{h,a} \,|\, Z_h] + \mathbb{E}[U_{h,a,a'} \,|\, Z_h] \\ &= \widetilde{\mathbb{F}}^\theta_{h,a'} \mathbb{E}[X_{h,a} \,|\, Z_h].\end{aligned} \tag{3.6}$$

We can assign any value $o \in \mathcal{O}$ to $Z_h = O_{h-1}$ in the condition of (3.6). Then we obtain

$$\begin{aligned}&\mathbb{P}^{\overline{\pi}}_{\mathrm{do}(A_{h-1:h}=(a,a'))}(O_{h:h+1} = \cdot \,|\, O_{h-1} = o) \\ &\quad = \widetilde{\mathbb{F}}^\theta_{h,a'} \mathbb{P}^{\overline{\pi}}_{\mathrm{do}(A_{h-1:h}=(a,a'))}(O_h = \cdot \,|\, O_{h-1} = o).\end{aligned} \tag{3.7}$$

For notational simplicity, we define the following two distributions

$$\begin{aligned}&P^\dagger_{h,a,a'}(o,o',o'') \\ &\qquad = \mathbb{P}^{\overline{\pi}}_{\mathrm{do}(A_{h-1:h}=(a,a'))}(O_{h-1:h+1} = (o,o',o'')),\end{aligned} \tag{3.8}$$

$$\begin{aligned}&P^\ddagger_{h,a}(o,o') \\ &\qquad = \mathbb{P}^{\overline{\pi}}_{\mathrm{do}(A_{h-1}=a)}(O_{h-1:h} = (o,o')).\end{aligned} \tag{3.9}$$

Then, we multiply both sides of (3.7) by $\mathbb{P}^{\overline{\pi}}(O_{h-1} = o_{h-1})$, which reweights the equations in (3.7) with the observation visitation measure induced by the behavior policy $\overline{\pi}$. We have the following moment equation for all $(h,a,a',o,o',o'') \in [H] \times \mathcal{A}^2 \times \mathcal{O}^3$,

$$P^\dagger_{h,a,a'}(o,o',o'') = (\mathbb{F}^\theta_{h,a'} P^\ddagger_{h,a})(o,o',o''), \tag{3.10}$$

Algorithm 1 General Framework

```
Require: Dataset $\mathcal{D}$, model parameter class $\Theta$, policy class $\Pi$, threshold parameter $\epsilon$ and oracle $\mathcal{J}$.
1: $(\widehat{P}^\dagger_{h,a,a'}, \widehat{P}^\ddagger_{h,a}) \leftarrow \mathsf{DistEst}(\mathcal{D})$.
2: $\widehat{\Theta} \leftarrow \{\theta \in \Theta : \sup_{(h,a,a')\in[H]\times\mathcal{A}^2}\{\|\mathbb{F}^\theta_{h,a'}\widehat{P}^\ddagger_{h,a} - \widehat{P}^\dagger_{h,a,a'}\|_1\} < \epsilon\}$.
3: $(\widehat{\pi}, \widehat{\theta}) = \mathrm{argmax}_{\pi\in\Pi}\,\mathrm{argmin}_{\theta\in\widehat{\Theta}}\,\mathcal{J}(\theta,\pi)$.
Ensure: $\widehat{\pi}, \widehat{\theta}$.
```

Algorithm 2 DistEst

```
Require: Dataset $\mathcal{D}_N = \{(s^n_h, a^n_h, \rho^n_h)\}^{N,H}_{n,h=1}$.
1: for $(h,a,a') \in \{2,\ldots,H\}\times\mathcal{A}^2$ do
2:     $\mathcal{D}_{h,a,a'} = \varnothing$
3:     for $n \in [N]$ do
4:         if $a^n_{h-1:h} = (a,a')$ then
5:             $\rho \leftarrow \rho^n_{h-1}\cdot\rho^n_h$
6:             $\mathcal{D}_{h,a,a'} \leftarrow \mathcal{D}_{h,a,a'} \cup \{o^n_{h-1:h+1}, \rho\}$
7:         end if
8:     end for
9:     Construct matrix $\Upsilon$ and vector $U$ as
           $[\Upsilon]_{i,j} = \langle\mathbb{K}\widetilde{\varphi}_i, \mathbb{K}\widetilde{\varphi}_j\rangle_{\mathcal{H}},$   (3.11)
           $[U]_i = \langle\mathbb{K}\widetilde{\varphi}_i, \widehat{\mathbb{K}}\mathcal{D}_{h,a,a'}\rangle_{\mathcal{H}},$   (3.12)
       for any $i,j\in[d]$.
10:    Construct estimators
           $\widehat{P}^\dagger_{h,a,a'}(o,o',o'') \leftarrow \widetilde{\varphi}(o,o',o'')^\top\Upsilon^{-1}U,$   (3.13)
           $\widehat{P}^\ddagger_{h,a}(o,o') \leftarrow \int_{\mathcal{O}}\widehat{P}^\dagger_{h,a,a'}(o,o',o'')\,\mathrm{d}o''.$   (3.14)
11: end for
Ensure: $P^\dagger_{h,a,a'}, P^\ddagger_{h,a}$.
```

where the linear operator $\mathbb{F}^\theta_{h,a}$ is adapted from $\widetilde{\mathbb{F}}^\theta_{h,a}$ such that

$$(\mathbb{F}^\theta_{h,a'}P^\ddagger_{h,a})(o,o',o'') = \int_{\mathcal{O}} P^\ddagger_{h,a}(o,\widetilde{o})\cdot\mathcal{F}^\theta_{h,a'}(\widetilde{o},o',o'')\,\mathrm{d}\widetilde{o},$$

where the integral kernel $\mathcal{F}^\theta_{h,a}$ is defined in (3.3).

Finally, to estimate the model parameter θ, we construct from the dataset empirical estimators of $P^\dagger_{h,a,a'}$ and $P^\ddagger_{h,a}$, denoted by $\widehat{P}^\dagger_{h,a,a'}$ and $\widehat{P}^\ddagger_{h,a}$, correspondingly. And we acquire a confidence region of θ by selecting θ such that $\widehat{P}^\dagger_{h,a,a'}$ and $\mathbb{F}^\theta_{h,a}\widehat{P}^\ddagger_{h,a}$ are close enough for all $(h,a,a')\in[H]\times\mathcal{A}^2$. Our general framework is summarized in Algorithm 1 We also introduce one way of estimating the empirical distributions via RKHS (Smola et al., 2007) in §3.2.

3.2. Distribution Estimators

In this section, we construct estimators for the distributions $P^\dagger_{h,a,a'}$ and $P^\ddagger_{h,a}$ from the dataset. Our method is summarized in Algorithm 2. We first construct empirical distributions of $P^\dagger_{h,a,a'}$ and $P^\ddagger_{h,a}$. This is done by decomposing the original dataset into trajectory pieces and grouping them based on the actions. See Line 2-7 of Algorithm 2. It's noteworthy that in $P^\dagger_{h,a,a'}$ and $P^\ddagger_{h,a}$, the actions a_{h-1} and a_h are specified regardless of the observations with the help of the do-operations. In other words, the corresponding distributions are induced by first executing the behavior policy $\overline{\pi}$ in the first $h-2$ steps, and then switch to a policy that selects a_{h-1} and a_h with probability 1 at step $h-1$ and h, respectively. However, the dataset is collected by policy $\overline{\pi}$ throughout, which incurs a distribution shift. To address that, we assume that the dataset contains an importance weight ρ for each (s,a) pair that equals the probability of the behavior policy $\overline{\pi}$ to take action a at that time. We can also estimate this importance weight with samples from the dataset instead of assuming its existence.

We then apply the reproducing kernel Hilbert space (RKHS) embedding (Smola et al., 2007) to estimate $P^\dagger_{h,a,a'}$ from $\mathcal{D}_{h,a,a'}$. Let $\mathcal{H}$ be an RKHS with the kernel $\mathcal{K}$ over $\mathcal{O}^3$, for example, the radial basis function (RBF) kernels (Smola & Schölkopf, 1998). With slight abuse of notation, we define the embedding operator $\mathbb{K}$ that embed both the probability distribution and the empirical distribution into the RKHS $\mathcal{H}$ as follows

$$(\mathbb{K}p)(x) = \int_{\mathcal{O}^3}\mathcal{K}(x',x)\cdot p(x')\,\mathrm{d}x', \tag{3.15}$$

$$(\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'})(x) = \sum_{(x',\rho_{x'})\in\mathcal{D}_{h,a,a'}}\frac{\mathcal{K}(x',x)}{N\rho_{x'}} \tag{3.16}$$

for any $x\in\mathcal{O}^3$. Then, we construct estimator $\widehat{P}^\dagger_{h,a,a'}$ to be the distribution that is the closest to $\mathcal{D}_{h,a,a'}$ when both are embedded into the RKHS $\mathcal{H}$. To narrow down the search range, we present the following lemma, which gives the linearity preserved in $P^\dagger_{h,a,a'}$ by Assumption 2.1.

Lemma 3.2. *Under Assumption 2.1, for any $(h,a,a')\in[H]\times\mathcal{A}^2$, there exists a feature mapping $\widetilde{\varphi}:\mathcal{O}^3\to\mathbb{R}^{d^3_\varphi}$ constructed from φ such that*

$$P^\dagger_{h,a,a'} \in \mathrm{linspan}(\widetilde{\varphi}),$$

where we considered $\widetilde{\varphi}$ as a set of functions and $\mathrm{linspan}(\widetilde{\varphi})$ *is the linear span of $\widetilde{\varphi}$.*

Proof. See §A.2 for a detailed proof. □

With Lemma 3.2, we no longer need to search over the whole space of $\Delta(\mathcal{O}^3)$, but only $\mathrm{linspan}(\widetilde{\varphi})$. That is to say, we let

$$\widehat{P}^{\dagger}_{h,a,a'} = \operatorname*{argmin}_{p \in \mathrm{linspan}(\widetilde{\varphi})} \|\mathbb{K}p - \mathbb{K}\mathcal{D}_{h,a,a'}\|_{\mathcal{H}}. \tag{3.17}$$

Since $\widetilde{\varphi}$ is constructed from φ, the minimization problem in (3.17) is tractable. We can also write $\widehat{P}^{\dagger}_{h,a,a'}(x) = \widetilde{\varphi}(x)^{\top}\widehat{w}_{h,a,a'}$, where $\widehat{w}$ is acquired by solving the following minimization problem

$$\widehat{w}_{h,a,a'} = \operatorname*{argmin}_{w \in \mathbb{R}^{d_{\varphi}^{3}}} \|(\mathbb{K}\widetilde{\varphi})^{\top} w - \mathbb{K}\mathcal{D}_{h,a,a'}\|_{\mathcal{H}}. \tag{3.18}$$

We further rewrite the objective function (3.18) as follows

$$\begin{aligned} &\|(\mathbb{K}\widetilde{\varphi})^{\top} w - \mathbb{K}\mathcal{D}_{h,a,a'}\|^{2}_{\mathcal{H}} \\ &\quad = \langle(\mathbb{K}\widetilde{\varphi})^{\top} w, (\mathbb{K}\widetilde{\varphi})^{\top} w\rangle_{\mathcal{H}} + \|\mathbb{K}\mathcal{D}_{h,a,a'}\|^{2}_{\mathcal{H}} \\ &\qquad - 2\langle(\mathbb{K}\widetilde{\varphi})^{\top} w, \mathbb{K}\mathcal{D}_{h,a,a'}\rangle_{\mathcal{H}} \\ &\quad = w^{\top}\Upsilon w + \|\mathbb{K}\mathcal{D}_{h,a,a'}\|^{2}_{\mathcal{H}} - 2U^{\top} w, \end{aligned} \tag{3.19}$$

where the matrix Υ and vector U are given in (3.11) and (3.12). The minimization problem of the quadratic form in (3.19) has the following closed form

$$\widehat{\omega}_{h,a,a'} = \Upsilon^{-1}U,$$

where the matrix Υ and vector U are defined in (3.11) and (3.12), respectively. Thus, we have the closed form for $P^{\dagger}$ presented in (3.13) of Algorithm 2. The construction of $\widehat{P}^{\ddagger}$ in (3.14) simply follows from the fact that

$$P^{\ddagger}_{h,a}(o,o') = \int_{\mathcal{O}} P^{\dagger}_{h,a,a'}(o,o',o'')\,\mathrm{d}o''.$$

4. Theory

We begin this section by introducing a useful analysis tool called the back operator in §4.1. Then, we introduce our main theoretical results in §4.2.

4.1. Backward Operator

The main challenge of analyzing a reinforcement learning algorithm on a POMDP is that we have to treat the whole history including all the previous observations and the actions as the "state" to maintain the Markov property since we do not have access to the true state. The troublesome dependency on the history makes it difficult to perform policy evaluation or value iteration like in the MDP cases (Sutton & Barto, 2018; Jin et al., 2020b; Yang & Wang, 2020; Wang et al., 2020a; Jin et al., 2021). We address this challenge by defining a Bellman operator like in MDPs, which maps the value function at step $h+1$ to its h-step correspondence. Such a process is also referred to as dynamic programming (Sutton & Barto, 2018). Recall that we exploit the undercompleteness assumption and the independence of observations given states when defining the forward operator $\mathbb{F}$ in (3.2). Following a similar idea, we define

$$\begin{aligned} \mathbb{B}^{\theta,\pi}_{h} V_{h+1}(\tau^{*}_{h}) = \int_{\mathcal{O}^{2}\times\mathcal{A}} & V_{h+1}(\tau^{*}_{h-1}, a^{*}_{h-1}, o_h, a_h, o_{h+1}) \\ \cdot\, \mathcal{F}^{\theta}_{h,a_h}(o^{*}_{h}, o_h, o_{h+1}) & \cdot \pi(a_h \mid \tau^{*}_{h-1}, a^{*}_{h-1}, o_h)\,\mathrm{d}a_h \mathrm{d}o_h \mathrm{d}o_{h+1} \end{aligned} \tag{4.1}$$

for any $\theta \in \Theta$, $\pi \in \Pi$, $h \in [H]$, and bounded function $V_{h+1} : \Gamma_{h+1} \to \mathbb{R}$. Here, the function $\mathcal{F}$ is defined in (3.3), and to avoid confusion, we write

$$\begin{aligned} \tau^{*}_{h} &= (\tau^{*}_{h-1}, a^{*}_{h-1}, o^{*}_{h}) = (\tau^{*}_{h-2}, a^{*}_{h-2}, o^{*}_{h-1}, a^{*}_{h-1}, o^{*}_{h}) \\ &= \cdots = (o^{*}_{1}, a^{*}_{1}, o^{*}_{2}, \ldots, a^{*}_{h-1}, o^{*}_{h}), \end{aligned}$$

for any $\tau^{*}_{h} \in \Gamma_h$. The construction of the forward operator is inspired by the Observable Operator Model (OOM) (Jaeger, 2000; Jin et al., 2020a) in learning latent variables. Those models are restricted to more strict settings, but we apply the ideas to our linear POMDPs. We then give the following lemma, which draws connection from our operator $\mathbb{B}^{\theta,\pi}_{h}$ to the Bellman operator in MDPs.

Lemma 4.1. *Under Assumption 2.1, 2.2, it holds for any $h \in [H]$, $\theta \in \Theta$, $\pi \in \Pi$, $\tau^{*}_{h-1} \in \Gamma_{h-1}$, $a^{*}_{h-1} \in \mathcal{A}$, $s^{*}_{h-1} \in \mathcal{S}$, and bounded function $V_{h+1} : \Gamma_{h+1} \to \mathbb{R}$ that*

$$\begin{aligned} &\mathbb{E}_{\theta,\pi}\big[\mathbb{B}^{\theta,\pi}_{h} V_{h+1}(\tau^{*}_{h-1}, a^{*}_{h-1}, o_h) \,\big|\, s^{*}_{h-1}, \tau^{*}_{h-1}, a^{*}_{h-1}\big] \\ &\quad = \mathbb{E}_{\theta,\pi}\big[V_{h+1}(\tau^{*}_{h-1}, a^{*}_{h-1}, o_h, a_h, o_{h+1}) \,\big|\, s^{*}_{h-1}, \tau^{*}_{h-1}, a^{*}_{h-1}\big], \end{aligned}$$

where the expectation is taken with respect to $o_h \sim \mathbb{P}^{\theta}_{h}(\cdot \mid s^{*}_{h-1}, a^{*}_{h-1})$, $a_h \sim \pi(\cdot \mid \tau^{*}_{h-1}, a^{*}_{h-1}, o_h)$, and $o_{h+1} \sim \mathbb{P}_{\theta}(\cdot \mid s^{*}_{h-1}, a^{*}_{h-1}, a_h)$.

Proof. See §B.1 for a detailed proof. □

Lemma 4.1 states that the operator $\mathbb{B}^{\theta,\pi}_{h}$ defined in (4.1) maps a function on the $(h+1)$-step history space to its h-step correspondence, which is the same role played by the Bellman operator in an MDP. We call $\mathbb{B}^{\theta,\pi}_{h}$ backward operator to distinguish from the forward operator $\mathbb{F}^{\theta}_{h}$. Both the definitions of the forward operator and backward operator take advantage of the undercompleteness assumption that enables us to retrieve information of the latent state from the distribution of the observations, which further explains why the undercompleteness assumption is the foundation of our algorithm and analysis.

To exploit the connection between the backward operator and the Bellman operator in MDPs, we define the following value function

$$V^{\theta,\pi}_{h} = \mathbb{B}^{\theta,\pi}_{h} \cdots \mathbb{B}^{\theta,\pi}_{H} R, \tag{4.2}$$

for any $h \in [H+1]$, $\theta \in \Theta$, and $\pi \in \Pi$. By definition, we let $V_{H+1}^{\theta,\pi} = R$. Here, the function $R : \Gamma_{H+1} \to \mathbb{R}$ calculates the total reward given a complete episode of observations and actions, that is,

$$R(\tau_{H+1}) = \sum_{h\in[H]} r(o_h, a_h), \tag{4.3}$$

for any $\tau_{H+1} \in \Gamma_{H+1}$. We give the following lemma, which reveals an integral form of $V_h^{\theta,\pi}$ and its connection with the total reward,

Lemma 4.2. *Under the same conditions as Lemma 4.1, for any $(\theta, \pi, h) \in \Theta \times \Pi \times [H]$ and $\tau_h^* = (\tau_{h-1}^*, a_{h-1}^*, o_h^*) \in \Gamma_h$, we have*

$$V_h^{\theta,\pi}(\tau_h^*) = \int_{\mathcal{S}} \mathbb{E}_{\theta,\pi}\Big[\sum_{h'\in[H]} r(o_{h'}, a_{h'}) \,\Big|\, s_h, \tau_{h-1}^*, a_{h-1}^*\Big] \cdot \xi_h^\theta(s_h, o_h^*)\,\mathrm{d}s_h. \tag{4.4}$$

Proof. See §B.2 for a detailed proof. □

By substitute o_h^* in (4.4) from a fixed observation to a random variable and taking expectation with respect to it, we can cancel off the function ξ on the right-hand side of (4.4), as stated in the following lemma.

Lemma 4.3. *Under the same conditions as Lemma 4.1, for any $(\theta, \pi, h) \in \Theta \times \Pi \times [H]$ and $(s_{h-1}^*, \tau_{h-1}^*, a_{h-1}^*) \in \mathcal{S} \times \Gamma_h \times \mathcal{A}$, we have*

$$\mathbb{E}_\theta\big[V_h^{\theta,\pi}(\tau_{h-1}^*, a_{h-1}^*, o_h) \,\big|\, s_{h-1}^*, \tau_{h-1}^*, a_{h-1}^*\big] = \mathbb{E}_{\theta,\pi}\Big[\sum_{h'\in[H]} r(o_{h'}, a_{h'}) \,\Big|\, s_{h-1}^*, \tau_{h-1}^*, a_{h-1}^*\Big]. \tag{4.5}$$

where the expectation on the left-hand side is taken with respect to $o_h \sim \mathbb{P}_\theta(\cdot \,|\, s_{h-1}^*, a_{h-1}^*)$.

Proof. See §B.3 for a detailed proof. □

Lemma 4.3 shows that the value function $V_h^{\theta,\pi}$ defined in (4.2) corresponds to the expected total reward following policy π from step h, just like in MDPs. By Lemma 4.3, we have $\mathcal{J}(\theta, \pi) = \mathbb{E}_\theta[V_1^{\theta,\pi}(o_1)]$ for $\mathcal{J}$ defined in (2.1) by setting h to 1.

4.2. Main Theorem

We begin this section by making more assumptions. Then, we present our main theoretical result in Theorem 4.6.

To start with, we note that the quality of the dataset is of great importance for offline RL algorithms (Wang et al., 2020b; Szepesvári, 2010). To avoid insufficient coverage of the offline dataset, we make the following assumption.

Assumption 4.4 (Coverage)**.** There exist absolute constants $\underline{C}, \overline{C} > 0$ such that (i). $|\mathcal{D}_{h,a,a'}| \ge N\underline{C}$ and (ii). $\mu_h^{\pi^*}(s)/\mu_h^{\bar\pi}(s) \le \overline{C}$ for any $(h, a, a', s) \in [H] \times \mathcal{A}^2 \times \mathcal{S}$.

Recall that $\mathcal{D}_{h,a,a'}$ is the dataset constructed by Algorithm 2, N is the cardinality of the offline dataset $\mathcal{D}_N$, function μ_h^π is the state distribution at step h induced by policy π, and π^* and $\bar\pi$ are the optimal policy and the behavior policy, respectively. Assumption 4.4 defines requirements for the offline dataset and the behavior policy that collects it. The condition (i) can be satisfied with high probability by a fixed behavior policy that explores all the actions with positive probabilities at any states. The condition (ii) states that the state distribution deviation of the behavior policy from the optimal policy should be bounded, which is a common assumption in the offline RL literature (Jin et al., 2020a; Agarwal et al., 2020; Liu et al., 2019b).

Recall that we apply RKHS embedding to estimate the distributions in §3.2 without specifying the kernel we are using. We keep the flexibility in the choices of the kernel but making the following assumption to specify the requirements for it.

Assumption 4.5 (Well-defined Kernel)**.** The kernel $\mathcal{K}$ is bounded, continuous, and positive definite. In particular, for any $x, x' \in \mathcal{O}^3$, we have $|\mathcal{K}(x, x')| \le 1$, and there exists a constant $\alpha > 0$ such that $\sigma_{\min}(\Upsilon) > \alpha$ where the matrix Υ is defined in (3.11).

Assumption 4.5 is a common assumption about the RKHS kernel, and can be satisfied by, for example, the radial basis function (RBF) kernels (Smola & Schölkopf, 1998). Now we're ready to present our main theorem.

Theorem 4.6. *Under Assumptions 2.1, 2.2, 4.4, 4.5, for any $\delta > 0$, if we let $\beta \ge (\gamma + 1) \cdot \beta_0$ where*

$$\beta_0 = \alpha^{-1} \cdot \sqrt{10d \cdot \log(2H|\mathcal{A}|^2/\delta)/\underline{C}}, \tag{4.6}$$

then, it holds with probability at least $1 - \delta$ that

$$\textit{Regret}(\widehat{\pi}) \le 2H^2|\mathcal{A}|^2\overline{C}\gamma^2\beta \cdot N^{-1/2},$$

where N is the size of the dataset.

Proof. See §5 for a sketched proof and §B.4 for a detailed proof. □

Theorem 4.6 guarantees an ϵ-optimal policy provided an offline dataset with sufficient coverage containing $\mathrm{poly}(H, |\mathcal{A}|, d_\phi, d_\psi) \cdot \widetilde{\mathcal{O}}(1/\epsilon^2)$ episodes, where H is the episode length, $|\mathcal{A}|$ is the cardinality of the action space, d_ψ and d_ψ are feature dimensions, and $\widetilde{\mathcal{O}}$ hides constant factors and logarithms. Such a rate is consistent with offline RL algorithms in MDPs such as Kidambi et al. (2020); Jin et al. (2021); Zanette & Brunskill (2019) under sufficient coverage. To our best knowledge, our algorithm is the first provably sample efficient offline RL algorithm for POMDPs that is not tabular.

5. Proof of Theorem 4.6

We begin our proof of the main theorem by presenting the following lemma.

Lemma 5.1. *Under the same conditions as Theorem 4.6, for any $\delta \in (0,1)$, the event E that*

$$\|\widehat{P}^{\ddagger}_{h,a} - P^{\ddagger}_{h,a}\|_1 \le \beta_0 \cdot N^{-1/2}, \qquad (5.1)$$

$$\|\widehat{P}^{\dagger}_{h,a,a'} - P^{\dagger}_{h,a,a'}\|_1 \le \beta_0 \cdot N^{-1/2}, \qquad (5.2)$$

for all $(h,a,a') \in [H] \times \mathcal{A}^2$ holds with probability at least $1-\delta$.

Proof. See §C.1 for a detailed proof. □

Lemma 5.1 gives our concentration analysis results for Algorithm 2. It states that our estimator $\widehat{P}^{\dagger}$ is close to the true distribution $P^{\dagger}$ in the sense of ℓ_1 norm. Under the event E defined above, it can be proved that the true model parameter lies in the confidence region constructed in Algorithm 2, as stated in the following lemma.

Lemma 5.2. *Under the event E defined in Lemma 5.1, it holds that $\theta^* \in \widehat{\Theta}$, where $\widehat{\Theta}$ is the output of Algorithm 2.*

Proof. See §C.2 for a detailed proof. □

By Lemma 5.2 and our pessimism principle in (3.1), we have

$$\mathcal{J}(\widehat{\theta}, \pi^*) \le \mathcal{J}(\widehat{\theta}, \widehat{\pi}) \le \mathcal{J}(\theta^*, \widehat{\pi}), \qquad (5.3)$$

under event E. As a consequence, we transform the regret defined in (2.3) as follows,

$$\begin{aligned}\text{Regret}(\widehat{\pi}) &= \mathcal{J}(\theta^*, \pi^*) - \mathcal{J}(\theta^*, \widehat{\pi}) \\ &\le \mathcal{J}(\theta^*, \pi^*) - \mathcal{J}(\widehat{\theta}, \pi^*),\end{aligned} \qquad (5.4)$$

The right-hand side of (5.4) is the estimation error of θ^* under the optimal policy π^*. Next, we adopt our backward operator to analysis such an estimation error. We define $\Delta\mathbb{B}_h^{(\theta,\theta'),\pi} V$ as the difference between $\mathbb{B}_h^{\theta,\pi}$ and $\mathbb{B}_h^{\theta',\pi}$ when applied to function $V : \Gamma_{h+1}$. That is,

$$\Delta\mathbb{B}_h^{(\theta,\theta'),\pi} V = \mathbb{B}_h^{\theta,\pi} V - \mathbb{B}_h^{\theta',\pi} V, \qquad (5.5)$$

for any $\theta, \theta' \in \Theta$, $\pi \in \Pi$, $h \in [H]$, and any bounded function $V : \Gamma_{h+1} \to \mathbb{R}$. With this notation, we give the following lemma, which further quantifies the right-hand side of (5.4).

Lemma 5.3 (Step-wise Error)**.** *Under Assumptions 2.1 and 2.2, it holds for any policy $\pi \in \Pi$ and $\theta, \theta' \in \Theta$ that*

$$\begin{aligned}&\mathcal{J}(\theta, \pi) - \mathcal{J}(\theta', \pi) \\ &\quad = \sum_{h \in [H]} \mathbb{E}_{\theta,\pi}\big[(\Delta\mathbb{B}_h^{(\theta,\theta'),\pi} V_{h+1}^{\theta',\pi})(\tau_h)\big],\end{aligned} \qquad (5.6)$$

where the function $V_{h+1}^{\theta,\pi}$ is defined in (4.2).

Proof. See §C.3 for a detailed proof. □

The right-hand side of (5.6) is a step-wise decomposition of the model estimation error. By conditioning the expectation on s_{h-1} and plugging it to (5.4), we obtain

$$\text{Regret}(\widehat{\pi}) \le \sum_{h \in [H]} \mathbb{E}_{\theta^*,\pi^*}\big[\epsilon_h(s_{h-1})\big], \qquad (5.7)$$

where we define the state-dependent error

$$\epsilon_h(s_{h-1}) = \big|\mathbb{E}_{\theta^*,\pi^*}\big[(\Delta\mathbb{B}_h^{(\theta^*,\widehat{\theta}),\pi^*} V_{h+1}^{\widehat{\theta},\pi^*})(\tau_h) \,\big|\, s_{h-1}\big]\big|. \qquad (5.8)$$

The dependence of the error on s_{h-1} weakens the impact of the optimal policy π^* on the trajectory τ_h. We manage to upper-bound the error under the behavior policy as follows.

Lemma 5.4. *For any $\delta \in (0,1)$, under the same conditions as Theorem 4.6, it holds with probability at least $1-\delta$ that*

$$\mathbb{E}_{\theta^*,\overline{\pi}}[\epsilon_h(s_{h-1})] \le 2H|\mathcal{A}|^2\gamma^2\beta \cdot N^{-1/2}, \qquad (5.9)$$

for any $h \in [H]$.

Proof. See §C.4 for a detailed proof. □

Notice that the expectation in (5.9) is taken with respect to the behavior policy, while the expectation in our goal in (5.7) is taken with respect to the optimal policy. To fill the gap, we invoke the second statement in Assumption 4.4 to change the measure from $\mu^{\overline{\pi}}$ to μ^{π^*}, which adds another factor $\overline{C}$ to our regret. See §B.4 for a detailed proof.

References

Agarwal, R., Schuurmans, D., and Norouzi, M. An optimistic perspective on offline reinforcement learning. In *International Conference on Machine Learning*, pp. 104–114. PMLR, 2020.

Anandkumar, A., Hsu, D., and Kakade, S. M. A method of moments for mixture models and hidden markov models. In *Conference on Learning Theory*, pp. 33–1. JMLR Workshop and Conference Proceedings, 2012.

Anandkumar, A., Ge, R., Hsu, D., Kakade, S. M., and Telgarsky, M. Tensor decompositions for learning latent variable models. *Journal of machine learning research*, 15:2773–2832, 2014.

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, 2020.

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

Bennett, A. and Kallus, N. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *arXiv preprint arXiv:2110.15332*, 2021.

Buckman, J., Gelada, C., and Bellemare, M. G. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

Cai, Q., Yang, Z., Lee, J. D., and Wang, Z. Neural temporal-difference learning converges to global optima. *arXiv preprint arXiv:1905.10027*, 2019.

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.

Fan, J., Wang, Z., Xie, Y., and Yang, Z. A theoretical analysis of deep q-learning. In *Learning for Dynamics and Control*, pp. 486–489. PMLR, 2020.

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.

Jaeger, H. Observable operator models for discrete stochastic time series. *Neural computation*, 12(6):1371–1398, 2000.

Jin, C., Kakade, S. M., Krishnamurthy, A., and Liu, Q. Sample-efficient reinforcement learning of undercomplete pomdps. *arXiv preprint arXiv:2006.12484*, 2020a.

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

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

Kidambi, R., Rajeswaran, A., Netrapalli, P., and Joachims, T. Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951*, 2020.

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., and Pérez, P. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 2021.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *arXiv preprint arXiv:2006.04779*, 2020.

Kwon, J., Efroni, Y., Caramanis, C., and Mannor, S. Rl for latent mdps: Regret guarantees and a lower bound. *arXiv preprint arXiv:2102.04939*, 2021.

Liao, L., Fu, Z., Yang, Z., Wang, Y., Kolar, M., and Wang, Z. Instrumental variable value iteration for causal offline reinforcement learning. *arXiv preprint arXiv:2102.09907*, 2021.

Liu, B., Cai, Q., Yang, Z., and Wang, Z. Neural proximal/trust region policy optimization attains globally optimal policy. *arXiv preprint arXiv:1906.10306*, pp. 10564–10575, 2019a.

Liu, Y., Swaminathan, A., Agarwal, A., and Brunskill, E. Off-policy policy gradient with state distribution correction. *arXiv preprint arXiv:1904.08473*, 2019b.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533, 2015.

Pathak, D., Gandhi, D., and Gupta, A. Self-supervised exploration via disagreement. In *International conference on machine learning*, pp. 5062–5071. PMLR, 2019.

Pearl, J. *Causality*. Cambridge university press, 2009.

Shi, C., Uehara, M., and Jiang, N. A minimax learning approach to off-policy evaluation in partially observable markov decision processes. *arXiv preprint arXiv:2111.06784*, 2021a.

Shi, T., Chen, D., Chen, K., and Li, Z. Offline reinforcement learning for autonomous driving with safety and exploration enhancement. *arXiv preprint arXiv:2110.07067*, 2021b.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676):354–359, 2017.

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pp. 13–31. Springer, 2007.

Smola, A. J. and Schölkopf, B. *Learning with kernels*, volume 4. Citeseer, 1998.

Sonabend-W, A., Lu, J., Celi, L. A., Cai, T., and Szolovits, P. Expert-supervised reinforcement learning for offline policy learning and evaluation. *arXiv preprint arXiv:2006.13189*, 2020.

Sondik, E. J. *The optimal control of partially observable Markov processes*. Stanford University, 1971.

Spaan, M. T. Partially observable markov decision processes. In *Reinforcement Learning*, pp. 387–414. Springer, 2012.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT press, 2018.

Szepesvári, C. Algorithms for reinforcement learning. *Synthesis lectures on artificial intelligence and machine learning*, 4(1):1–103, 2010.

Tang, S. and Wiens, J. Model selection for offline reinforcement learning: Practical considerations for healthcare settings. In *Machine Learning for Healthcare Conference*, pp. 2–35. PMLR, 2021.

Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W., Dudzik, A., Huang, A., Georgiev, P., Powell, R., et al. Alphastar: Mastering the Real-Time Strategy Game StarCraft II, 2019.

Wang, L., Cai, Q., Yang, Z., and Wang, Z. Neural policy gradient methods: Global optimality and rates of convergence. *arXiv preprint arXiv:1909.01150*, 2019.

Wang, R., Du, S. S., Yang, L. F., and Salakhutdinov, R. On reward-free reinforcement learning with linear function approximation. *arXiv preprint arXiv:2006.11274*, 2020a.

Wang, R., Foster, D. P., and Kakade, S. M. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020b.

Wei, H., Zheng, G., Yao, H., and Li, Z. Intellilight: A reinforcement learning approach for intelligent traffic light control. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2496–2505, 2018.

Xiong, Y., Chen, N., Gao, X., and Zhou, X. Sublinear regret for learning pomdps. *arXiv preprint arXiv:2107.03635*, 2021.

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.

Ye, D., Liu, Z., Sun, M., Shi, B., Zhao, P., Wu, H., Yu, H., Yang, S., Wu, X., Guo, Q., et al. Mastering complex control in moba games with deep reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 6672–6679, 2020.

Yu, C., Liu, J., Nemati, S., and Yin, G. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

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

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

Zhou, D., He, J., and Gu, Q. Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*, pp. 12793–12802. PMLR, 2021.

A. Linear POMDP

A.1. Proof of Lemma 3.1

For any $(h, a, a') \in [H] \times \mathcal{A}^2$, it holds that

$$\begin{aligned}\mathbb{F}^{\theta^*}_{h,a'}\mathbb{E}[X_{h,a}\,|\,O_{1:h-1}](o,o') &= \int_{\mathcal{O}} \mathbb{P}^{\bar\pi}_{\mathrm{do}(A_{h-1}=a)}(O_h=\widetilde{o}\,|\,O_{1:h-1})\cdot\mathcal{F}^{\theta^*}_{h,a'}(\widetilde{o},o,o')\,\mathrm{d}\widetilde{o}\\ &= \int_{\mathcal{O}\times\mathcal{S}^4} \mathbb{P}^{\bar\pi}(S_{h-1}=s\,|\,O_{1:h-1})\cdot\mathcal{T}_{h-1}(s'\,|\,s,a)\cdot\mathcal{E}_h(\widetilde{o}\,|\,s')\\ &\qquad\cdot\xi_h(\widetilde{s},\widetilde{o})\cdot\mathcal{E}_h(o\,|\,\widetilde{s})\cdot\mathcal{T}_h(\widetilde{s}'\,|\,\widetilde{s},a')\cdot\mathcal{E}_{h+1}(o'\,|\,\widetilde{s}')\,\mathrm{d}s\,\mathrm{d}s'\,\mathrm{d}\widetilde{o}\,\mathrm{d}\widetilde{s}\,\mathrm{d}\widetilde{s}'.\end{aligned}$$

By Assumption 2.2, we have

$$\int_{\mathcal{S}\times\mathcal{O}} \xi_h(\widetilde{s},\widetilde{o})\cdot\mathcal{E}_h(\widetilde{o}\,|\,s')\cdot\mathcal{T}_{h-1}(s'\,|\,s,a)\,\mathrm{d}s'\,\mathrm{d}\widetilde{o} = \mathcal{T}_{h-1}(\widetilde{s}\,|\,s,a). \tag{A.1}$$

Thus,

$$\begin{aligned}&\mathbb{F}^{\theta^*}_{h,a'}\mathbb{E}[X_{h,a}\,|\,O_{1:h-1}](o,o')\\ &\quad= \int_{\mathcal{S}^3} \mathbb{P}^{\bar\pi}(S_{h-1}=s\,|\,O_{1:h-1})\cdot\mathcal{T}_{h-1}(\widetilde{s}\,|\,s,a)\cdot\mathcal{E}_h(o\,|\,\widetilde{s})\cdot\mathcal{T}_h(\widetilde{s}'\,|\,\widetilde{s},a')\cdot\mathcal{E}_{h+1}(o'\,|\,\widetilde{s}')\,\mathrm{d}s\,\mathrm{d}\widetilde{s}\,\mathrm{d}\widetilde{s}'\\ &\quad= \mathbb{P}^{\bar\pi}_{\mathrm{do}(A_{h-1:h}=(a,a'))}[O_{h:h+1}=(o,o')\,|\,O_{1:h-1}]\\ &\quad= \mathbb{E}[Y_{h,a,a'}\,|\,O_{1:h-1}](o,o').\end{aligned}$$

Then,

$$\begin{aligned}Y_{h,a,a'} &= \mathbb{F}^{\theta^*}_{h,a'}X_{h,a} + \mathbb{F}^{\theta^*}_{h,a'}(\mathbb{E}[X_{h,a}\,|\,O_{1:h-1}] - X_{h,a}) + Y_{h,a,a'} - \mathbb{E}[Y_{h,a,a'}\,|\,O_{1:h-1}] &\text{(A.2)}\\ &= \mathbb{F}^{\theta^*}_{h,a'}X_{h,a} + \mathbb{F}^{\theta^*}_{h,a'}(\mathbb{E}[X_{h,a}\,|\,O_{1:h-1}] - X_{h,a}) + U_{h,a,a'}. &\text{(A.3)}\end{aligned}$$

We define the following linear operator $\widetilde{\mathbb{L}} : (\mathcal{O}^2 \to \mathbb{R}) \to (\mathcal{O} \to \mathbb{R})$

$$(\widetilde{\mathbb{L}}f)(o) = \int_{\mathcal{O}} f(o,o')\,\mathrm{d}o'. \tag{A.4}$$

Then, we can write

$$Y_{h,a,a'} = \mathbb{F}^{\theta^*}_{h,a'}X_{h,a} + \mathbb{F}^{\theta^*}_{h,a'}\widetilde{\mathbb{L}}U_{h,a,a'} + U_{h,a,a'}.$$

Since both $\mathbb{F}^{\theta}_{h,a'}$ and $\mathbb{L}$ are linear operators, we conclude the proof of Lemma 3.1.

A.2. Additional Linearity

In this section, we show two additional linear structures inside our linear POMDP besides (2.4) and (2.5). We first prove Lemma 3.2, which states that the joint probability of consecutive three observations given necessary actions is linear in a constructed feature mapping. Then, we define an inverse transition function and show this function is also linear in a constructed feature mapping.

A.2.1. Linearity of $P^\dagger$

Proof of Lemma 3.2. We write down the explicit form of $P^\dagger$

$$
\begin{aligned}
&P^\dagger_{h,a_{h-1},a_h}(o_{h-1},o_h,o_{h+1}) \\
&\quad = \mathbb{P}_{\theta^*,\bar{\pi}}(o_{h-1},o_h,o_{h+1} \,|\, a_{h-1},a_h) \\
&\quad = \int_{\mathcal{S}^3} \mathcal{E}^{\theta^*}_{h-1}(o_{h-1} \,|\, s_{h-1}) \cdot \mathcal{T}^{\theta^*}_{h-1}(s_h \,|\, s_{h-1},a_{h-1}) \cdot \mathcal{E}^{\theta^*}_h(o_h \,|\, s_h) \cdot \mathcal{T}^{\theta^*}_h(s_{h+1} \,|\, s_h,a_h) \\
&\qquad\quad \cdot \mathcal{E}^{\theta^*}_{h+1}(o_{h+1} \,|\, s_{h+1}) \cdot \mu^{\bar{\pi}}_{h-1}(s_{h-1}) \,\mathrm{d}s_{h-1}\,\mathrm{d}s_h\,\mathrm{d}s_{h+1} \\
&\quad = \int_{\mathcal{S}^3} \varphi(o_{h-1})^\top \eta^{\theta^*}_{h-1}(s_{h-1}) \cdot \mathcal{T}^{\theta^*}_{h-1}(s_h \,|\, s_{h-1},a_{h-1}) \cdot \varphi(o_h)^\top \eta^{\theta^*}_h(s_h) \cdot \mathcal{T}^{\theta^*}_h(s_{h+1} \,|\, s_h,a_h) \\
&\qquad\quad \cdot \varphi(o_{h+1})^\top \eta^{\theta^*}_{h+1}(s_{h+1}) \cdot \mu^{\bar{\pi}}_h(s_{h-1}) \,\mathrm{d}s_{h-1}\,\mathrm{d}s_h\,\mathrm{d}s_{h+1},
\end{aligned}
$$

where the second equality follows from Assumption 2.1, and we exploit the fact that the observations are independent given corresponding observations and the actions in the conditions are decoupled from the history as explained in (3.8). Then, by rearranging the terms, we know there exists coefficient $w_{h,a,a',i,j,k}$ for any $(h,a,a',i,j,k) \in [H] \times \mathcal{A}^2 \times [d_\varphi]^3$ such that

$$
P^\dagger_{h,a,a'}(o,o',o'') = \sum_{(i,j,k)\in[d_\varphi]^3} \varphi_i(o)\varphi_j(o')\varphi_k(o'') \cdot w_{h,a,a',i,j,k}.
$$

Thus, by defining $\widetilde{\varphi}(o,o',o'') = \sum_{(i,j,k)\in[d_\varphi]^3} \varphi_i(o)\varphi_j(o')\varphi_k(o'')$, we conclude the proof of Lemma 3.2. □

A.2.2. Linearity of $\mathcal{I}$

In this section, we show the following inverse transition function is also linear in ψ

$$
\mathcal{I}^\theta_h(s_h \,|\, a_h,o_{h+1}) = \mathbb{P}_\theta(s_h \,|\, a_h,o_{h+1}), \quad \text{for any } (s_h,a_h,o_{h+1}) \in \mathcal{S}\times\mathcal{A}\times\mathcal{O}. \tag{A.5}
$$

Here, the action a_h in the conditions is also decoupled from the history, i.e., the value of it is selected by us regardless of the history and the policy, like in (3.8) and (3.9). Thus, the function $\mathcal{I}$ is a property of the Linear POMDP and does not involve any policy π. The function $\mathcal{I}$ defines the distribution of the state given the future action and observation. By Bayesian's theorem, we can write

$$
\mathcal{I}^\theta_h(s_h \,|\, a_h,o_{h+1}) = \frac{\mathbb{P}_\theta(o_{h+1} \,|\, s_h,a_h) \cdot \mathbb{P}_{\theta,\pi}(s_h)}{\mathbb{P}_{\theta,\pi}(o_{h+1} \,|\, a_h)}. \tag{A.6}
$$

Note that

$$
\begin{aligned}
\mathbb{P}_\theta(o_{h+1} \,|\, s_h,a_h) &= \int_{\mathcal{S}} \mathcal{T}^\theta_h(s_{h+1} \,|\, s_h,a_h) \cdot \mathcal{E}^\theta_h(o_{h+1} \,|\, s_{h+1}) \,\mathrm{d}s_{h+1} \\
&= \phi(s_h,a_h)^\top M^\theta_h \int_{\mathcal{S}} \psi(s_{h+1}) \cdot \mathcal{E}^\theta_h(o_{h+1} \,|\, s_{h+1}) \,\mathrm{d}s_{h+1},
\end{aligned} \tag{A.7}
$$

and

$$
\mathbb{P}_{\theta,\pi}(s_h) = \mathbb{E}_{\theta,\pi}\big[\mathcal{T}^\theta_{h-1}(s_h \,|\, s_{h-1},a_{h-1})\big] = \mathbb{E}_{\theta,\pi}\big[\phi(s_{h-1},a_{h-1})^\top\big] M^\theta_h \psi(s_h). \tag{A.8}
$$

Plugging (A.8) and (A.7) into (A.6), we obtain

$$
\mathcal{I}^\theta_h(s_h \,|\, a_h,o_{h+1}) = \phi(s_h,a_h)^\top M' \psi(s_h) = \sum_{(i,j)\in[d_\phi]\times[d_\psi]} \phi_i(s_h,a_h)\psi_j(s_h) M'_{i,j}, \tag{A.9}
$$

if we define

$$
M' = \frac{M^\theta_h \big(\int_{\mathcal{S}} \psi(s_{h+1}) \cdot \mathcal{E}^\theta_h(o_{h+1} \,|\, s_{h+1}) \,\mathrm{d}s_{h+1}\big) \mathbb{E}_{\theta,\pi}\big[\phi(s_{h-1},a_{h-1})^\top\big] M^\theta_h}{\mathbb{P}_{\theta,\pi}(o_{h+1} \,|\, a_h)}.
$$

By (A.9), we know $\mathcal{I}_h^\theta(\cdot \,|\, a_h, o_{h+1})$ is a linear combination of the functions $\{\phi_i(\cdot, a_h)\cdot\psi_j(\cdot)\}_{(i,j)\in[d_\phi]\times[d_\psi]}$. Thus, by normalizing each function in it as probability distribution function and eliminating linearly dependent elements. We obtain a new feature mapping $\widetilde{\psi}: \mathcal{S}\to\mathbb{R}^{d_{\widetilde{\psi}}}$ such that $\mathcal{I}_h^\theta(\cdot \,|\, a_h, o_{h+1}) \in \mathrm{linspan}(\widetilde{\psi})$, and we have $d_{\widetilde{\psi}} = d_\psi(d_\phi+1)$. With the new feature mapping, we obtain

$$\mathcal{T}_h^\theta(\cdot \,|\, s_h, a_h) \in \mathrm{linspan}(\psi'), \quad \mathcal{I}_h^\theta(\cdot \,|\, a_h, o_{h+1}) \in \mathrm{linspan}(\psi').$$

Thus, we can always substitute ψ with this new feature mapping ψ' without violating the definition of the linear MDP. Throughout the appendix, we assume the property $\mathcal{I}_h^\theta(\cdot \,|\, a_h, o_{h+1}) \in \mathrm{linspan}(\psi)$ holds for the default ψ of the linear POMDP.

A.3. Sufficient Conditions for the Undercompleteness Assumption

In this section, we discuss the sufficient conditions for Assumption 2.2 to hold. We give the following lemma.

Lemma A.1. *Assumption 2.2 holds when the following two conditions hold.*

(a). There exists mapping $\overline{\phi}: \mathcal{O}\to\mathbb{R}^{d_\varphi}$ and constant γ_1 such that:

$$\int_{\mathcal{O}} \overline{\varphi}(o)\varphi(o)^\top \,\mathrm{d}o = I \quad \textit{and} \quad \sup_{o\in\mathcal{O}} \|\overline{\varphi}(o)\|_1 \le \gamma_1. \tag{A.10}$$

(b). For any $h\in[H]$ and $\theta\in\Theta$, there exists constant γ_2 such that

$$\|(\Psi_h^\theta)^+\|_{1\mapsto 1} \le \gamma_2, \quad \textit{where} \;\; \Psi_h^\theta = \int_{\mathcal{S}} \eta_h^\theta(s)\psi(s)^\top \,\mathrm{d}s \tag{A.11}$$

has full column rank, and $A^+ = (A^\top A)^{-1}A^\top$ is the Moore-Penrose inverse of matrix A.

Proof of Lemma A.1. We let

$$\xi_h^\theta(s,o) = \psi(s)^\top(\Psi_h^\theta)^+\overline{\varphi}(o).$$

We first prove $\mathbb{U}_h^\theta\mathbb{O}_h^\theta f = f$. For any $f\in\mathrm{linspan}(\psi)$ with the coefficient x_f, i.e., $f(\cdot) = \psi(\cdot)^\top x_f$, we have

$$\begin{aligned}\mathbb{U}_h^\theta\mathbb{O}_h^\theta f(s) &= \int_{\mathcal{S}\times\mathcal{O}} \xi_h^\theta(s,o)\cdot\mathcal{E}_h^\theta(o\,|\,s')\cdot\psi(s')^\top x_f \,\mathrm{d}s'\,\mathrm{d}o\\ &= \psi(s)^\top(\Psi_h^\theta)^+\biggl(\int_{\mathcal{S}\times\mathcal{O}} \overline{\varphi}(o)\cdot\varphi(o)^\top\eta_h^\theta(s')\cdot\psi(s')^\top \,\mathrm{d}s'\,\mathrm{d}o\biggr)x_f.\end{aligned}$$

By the definition of φ and Ψ_h^θ in (A.10) and (A.11), respectively, we have

$$\mathbb{U}_h^\theta\mathbb{O}_h^\theta f(s) = \psi(s)^\top(\Psi_h^\theta)^+\biggl(\int_{\mathcal{S}} \eta_h^\theta(s')\cdot\psi(s')^\top \,\mathrm{d}s'\biggr)x_f = \psi(s)^\top(\Psi_h^\theta)^+\Psi_h^\theta x_f = f(s).$$

Next, we prove $\|\mathbb{U}_h^\theta\|_{1\mapsto 1}$ is upper-bounded. For any $f\in L^1(\mathcal{O})$, $h\in[H]$ and $\theta\in\Theta$, we have

$$\begin{aligned}\|\mathbb{U}_h^\theta f\|_1 &= \int_{\mathcal{S}}\biggl|\int_{\mathcal{O}}\psi(s)^\top(\Psi_h^\theta)^+\overline{\varphi}(o)\cdot f(o)\,\mathrm{d}o\biggr|\,\mathrm{d}s\\ &\le \int_{\mathcal{S}}\psi(s)^\top\biggl|(\Psi_h^\theta)^+\int_{\mathcal{O}}\overline{\varphi}(o)\cdot f(o)\,\mathrm{d}o\biggr|\,\mathrm{d}s = \biggl\|(\Psi_h^\theta)^+\int_{\mathcal{O}}\overline{\varphi}(o)\cdot f(o)\,\mathrm{d}o\biggr\|_1,\end{aligned}$$

where the equality follows from the fact that $\psi(\cdot)$ is a probability distribution over $\mathcal{S}$. By the triangle inequality and (A.11), we have

$$\|\mathbb{U}_h^\theta f\|_1 = \biggl\|(\Psi_h^\theta)^+\int_{\mathcal{O}}\overline{\varphi}(o)\cdot f(o)\,\mathrm{d}o\biggr\|_1 \le \gamma_2\cdot\biggl\|\int_{\mathcal{O}}\overline{\varphi}(o)\cdot f(o)\,\mathrm{d}o\biggr\|_1 \le \gamma_2\cdot\gamma_1\|f\|_1.$$

Then, we conclude the proof by letting $\gamma = \gamma_1\gamma_2$. □

B. Missing Proofs in §4

B.1. Proof of Lemma 4.1

Proof of Lemma 4.1. By the definition of $\mathbb{B}_h^{\theta,\pi}$ in (4.1), for any $s_{h-1}^* \in \mathcal{S}$, $\tau_{h-1}^* \in \Gamma_{h-1}$, $a_{h-1}^* \in \mathcal{A}$ and bounded function $V_{h+1} : \Gamma_{h+1} \to \mathbb{R}$, we have

$$\begin{aligned}
&\mathbb{E}_{\theta,\pi}\big[(\mathbb{B}_h^{\theta,\pi} V)(\tau_{h-1}^*, a_{h-1}^*, o_h) \,\big|\, s_{h-1}^*, \tau_{h-1}^*, a_{h-1}^*\big] \\
&\quad = \int_{\mathcal{O}^3 \times \mathcal{A} \times \mathcal{S}} \pi(a_h \,|\, \tau_{h-1}^*, a_{h-1}^*, o_h') \cdot V_{h+1}(\tau_{h-1}^*, a_{h-1}^*, o_h', a_h, o_{h+1}) \\
&\qquad\qquad \cdot \mathbb{P}_\theta(o_h', o_{h+1} \,|\, s_h, a_h) \cdot \xi_h^\theta(s_h, o_h) \cdot \mathbb{P}_\theta(o_h \,|\, s_{h-1}^*, a_{h-1}^*)\, \mathrm{d}o_h\, \mathrm{d}o_h'\, \mathrm{d}o_{h+1}\, \mathrm{d}a_h.
\end{aligned} \tag{B.1}$$

The trick here is since we are taking expectation with respect to o_h conditional on s_{h-1}^* and a_{h-1}^*, the emission function $\mathcal{E}_h^\theta$ contained in the expectation will cancel off ξ_h^θ. To see that, we apply (**??**) which invokes Assumption 2.2. It holds that

$$\begin{aligned}
&\mathbb{E}_{\theta,\pi}\big[(\mathbb{B}_h^{\theta,\pi} V)(\tau_{h-1}^*, a_{h-1}^*, o_h) \,\big|\, s_{h-1}^*, \tau_{h-1}^*, a_{h-1}^*\big] \\
&\quad = \int_{\mathcal{O}^2 \times \mathcal{A} \times \mathcal{S}} \pi(a_h \,|\, \tau_{h-1}^*, a_{h-1}^*, o_h') \cdot V_{h+1}(\tau_{h-1}^*, a_{h-1}^*, o_h', a_h, o_{h+1}) \\
&\qquad\qquad \cdot \mathbb{P}_\theta(o_h', o_{h+1} \,|\, s_h, a_h) \cdot \mathcal{T}_h^\theta(s_h \,|\, s_{h-1}^*, a_{h-1}^*)\, \mathrm{d}o_h'\, \mathrm{d}o_{h+1}\, \mathrm{d}a_h \\
&\quad = \int_{\mathcal{O}^2 \times \mathcal{A}} V_{h+1}(\tau_{h-1}^*, a_{h-1}^*, o_h', a_h, o_{h+1}) \cdot \mathbb{P}_{\theta,\pi}(o_h', a_h, o_{h+1}, \,|\, \tau_{h-1}^*, s_{h-1}^*, a_{h-1}^*)\, \mathrm{d}o_h'\, \mathrm{d}a_h\, \mathrm{d}o_{h+1},
\end{aligned}$$

which concludes the proof. □

B.2. Proof of Lemma 4.2

Proof of Lemma 4.2. We prove the lemma by induction. For $h = H$, by the definition of $\mathbb{B}_H^{\theta,\pi}$ and $V_H^{\theta,\pi}$ in (4.1) and (4.2), respectively, and the additional definition $V_{H+1}^{\theta,\pi} = R$, for any $\tau_H^* = (\tau_{H-1}^*, a_{H-1}^*, o_H^*) \in \Gamma_H$, we have

$$\begin{aligned}
V_H^{\theta,\pi}(\tau_H^*) &= (\mathbb{B}_H^{\theta,\pi} V_{H+1}^{\theta,\pi})(\tau_H^*) \\
&= \int_{\mathcal{S} \times \mathcal{O}^2} \mathbb{E}_{a_H \sim \pi(\cdot \,|\, \tau_{H-1}^*, a_{H-1}^*, o_H)}\big[V_{H+1}^{\theta,\pi}(\tau_{H-1}^*, a_{H-1}^*, o_H, a_H, o_{H+1}) \cdot \mathbb{P}_\theta(o_H, o_{H+1} \,|\, s_H, a_H)\big] \\
&\qquad\qquad \cdot \xi_h^\theta(s_H, o_H^*)\, \mathrm{d}s_H\, \mathrm{d}o_H\, \mathrm{d}o_{H+1} \\
&= \int_{\mathcal{S} \times \mathcal{O}^2} \mathbb{E}_{a_H \sim \pi(\cdot \,|\, \tau_{H-1}^*, a_{H-1}^*, o_H)}\Big[\Big(r(o_H, a_H) + \sum_{h \in [H-1]} r(o_H^*, a_h^*)\Big) \cdot \mathbb{P}_\theta(o_H, o_{H+1} \,|\, s_H, a_H)\Big] \\
&\qquad\qquad \cdot \xi_h^\theta(s_H, o_H^*)\, \mathrm{d}s_H\, \mathrm{d}o_H\, \mathrm{d}o_{H+1} \\
&= \int_{\mathcal{S}} \Big(\mathbb{E}_{\theta,\pi}\big[r(o_H, a_H) \,\big|\, s_H, \tau_{H-1}^*, a_{H-1}^*\big] + \sum_{h \in [H-1]} r(o_H^*, a_h^*)\Big) \cdot \xi_h^\theta(s_H, o_H^*)\, \mathrm{d}s_H,
\end{aligned}$$

which satisfies the statement (4.4). Assume the lemma holds for the value function of the $(h+1)$-th step. Then, we have

$$\begin{aligned}
V_h^{\theta,\pi}(\tau_h) &= (\mathbb{B}_h^{\theta,\pi} V_{h+1}^{\theta,\pi})(\tau_h) \\
&= \int_{\mathcal{S} \times \mathcal{O}^2} \mathbb{E}_{a_h \sim \pi(\cdot \,|\, \tau_{H-1}^*, a_{H-1}^*, o_h)}\big[V_{h+1}^{\theta,\pi}(\tau_{H-1}^*, a_{H-1}^*, o_h, a_h, o_{h+1}) \cdot \mathbb{P}_\theta(o_h, o_{h+1} \,|\, s_h, a_h)\big] \\
&\qquad\qquad \cdot \xi_h^\theta(s_h, o_H^*)\, \mathrm{d}s_h\, \mathrm{d}o_h\, \mathrm{d}o_{h+1} \\
&= \int_{\mathcal{S}^2 \times \mathcal{O}^2} \mathbb{E}_{a_h \sim \pi(\cdot \,|\, \tau_{H-1}^*, a_{H-1}^*, o_h)}\Big[\mathbb{E}_{\theta,\pi}\Big[\sum_{i \in [H]} r(o_i, a_i) \,\Big|\, s_{h+1}, \tau_{H-1}^*, a_{H-1}^*, o_h, a_h\Big] \cdot \mathbb{P}_\theta(o_h, o_{h+1} \,|\, s_h, a_h)\Big] \\
&\qquad\qquad \cdot \xi_{h+1}^\theta(s_{h+1}, o_{h+1}) \cdot \xi_h^\theta(s_h, o_H^*)\, \mathrm{d}s_h\, \mathrm{d}s_{h+1}\, \mathrm{d}o_h\, \mathrm{d}o_{h+1}.
\end{aligned} \tag{B.2}$$

We write down the explicit form of $\mathbb{P}_\theta(o_h, o_{h+1} \,|\, s_h, a_h)$ as follows

$$\mathbb{P}_\theta(o_h, o_{h+1} \,|\, s_h, a_h) = \mathcal{E}_h^\theta(o_h \,|\, s_h) \cdot \int_{\mathcal{S}} \mathcal{E}_{h+1}^\theta(o_{h+1} \,|\, s_{h+1}') \cdot \mathcal{T}_h^\theta(s_{h+1}' \,|\, s_h, a_h)\, \mathrm{d}s_{h+1}', \tag{B.3}$$

and note that, following from Assumption 2.2, we have

$$\int_{\mathcal{S}\times\mathcal{O}} \xi^{\theta}_{h+1}(s_{h+1}, o_{h+1}) \cdot \mathcal{E}^{\theta}_{h+1}(o_{h+1} \,|\, s'_{h+1}) \cdot \mathcal{T}^{\theta}_{h}(s'_{h+1} \,|\, s_h, a_h)\, \mathrm{d}s'_{h+1}\, \mathrm{d}o_{h+1} = \mathcal{T}^{\theta}_{h}(s_{h+1} \,|\, s_h, a_h), \tag{B.4}$$

since $\mathcal{T}^{\theta}_{h}(\cdot \,|\, s_h, a_h)$ is in the linear span of ψ. Plugging (B.3) and (B.4) into (B.2), we have

$$\begin{aligned} V^{\theta,\pi}_{h}(\tau_h) &= \int_{\mathcal{S}^2\times\mathcal{O}} \mathbb{E}_{a_h\sim\pi(\cdot \,|\, \tau^*_{H-1}, a^*_{H-1}, o_h)}\Big[\mathbb{E}_{\theta,\pi}\Big[\sum_{i\in[H]} r(o_i, a_i) \,\Big|\, s_{h+1}, \tau^*_{H-1}, a^*_{H-1}, o_h, a_h\Big]\Big] \\ &\qquad\qquad \cdot \mathcal{E}^{\theta}_{h}(o_h \,|\, s_h) \cdot \mathcal{T}^{\theta}_{h}(s_{h+1} \,|\, s_h, a_h) \cdot \xi^{\theta}_{h}(s_h, o^*_H)\, \mathrm{d}s_h\, \mathrm{d}s_{h+1}\, \mathrm{d}o_h \\ &= \int_{\mathcal{S}} \mathbb{E}_{\theta,\pi}\Big[\sum_{i\in[H]} r(o_i, a_i) \,\Big|\, s_h, \tau^*_{H-1}, a^*_{H-1}\Big] \cdot \xi^{\theta}_{h}(s_h, o^*_H)\, \mathrm{d}s_h\, \mathrm{d}o_h, \end{aligned}$$

which concludes the proof of Lemma 4.2. □

B.3. Proof of Lemma 4.3

Proof of Lemma 4.3. By the definition of the value function in (4.2), for any $h \in [H]$, $s^*_{h-1} \in \mathcal{S}$, $\tau^*_{h-1} \in \Gamma_{h-1}$, and $a^*_{h-1} \in \mathcal{A}$, we have

$$\mathbb{E}_{\theta}\big[V^{\theta,\pi}_{h}(\tau^*_{h-1}, a^*_{h-1}, o_h) \,\big|\, s^*_{h-1}, \tau^*_{h-1}, a^*_{h-1}\big] = \mathbb{E}_{\theta}\big[(\mathbb{B}^{\theta,\pi}_{h} \cdots \mathbb{B}^{\theta,\pi}_{H} R)(\tau^*_{h-1}, a^*_{h-1}, o_h) \,\big|\, s^*_{h-1}, \tau^*_{h-1}, a^*_{h-1}\big].$$

By Lemma 4.1, we have

$$\begin{aligned} &\mathbb{E}_{\theta}\big[V^{\theta,\pi}_{h}(\tau^*_{h-1}, a^*_{h-1}, o_h) \,\big|\, s^*_{h-1}, \tau^*_{h-1}, a^*_{h-1}\big] \\ &\quad = \mathbb{E}_{\theta,\pi}\big[(\mathbb{B}^{\theta,\pi}_{h+1} \cdots \mathbb{B}^{\theta,\pi}_{H} R)(\tau^*_{h-1}, a^*_{h-1}, o_h, a_h, o_{h+1}) \,\big|\, s^*_{h-1}, \tau^*_{h-1}, a^*_{h-1}\big]. \end{aligned}$$

By the tower property and applying Lemma 4.1 repeatedly, we obtain

$$\begin{aligned} &\mathbb{E}_{\theta}\big[V^{\theta,\pi}_{h}(\tau^*_{h-1}, a^*_{h-1}, o_h) \,\big|\, s^*_{h-1}, \tau^*_{h-1}, a^*_{h-1}\big] \\ &\quad = \cdots = \mathbb{E}_{\theta,\pi}\big[R(\tau^*_{h-1}, a^*_{h-1}, o_h, a_h, o_{h+1}, \ldots, a_H, o_{H+1}) \,\big|\, s^*_{h-1}, \tau^*_{h-1}, a^*_{h-1}\big], \end{aligned}$$

which concludes the proof. □

B.4. Proof of the Main Theorem

Proof of Theorem 4.6. We condition our proof on the event E which holds with probability at least $1-\delta$. By Lemma 5.2, we have $\theta^* \in \widehat{\Theta}$. Thus, following from (3.1), we have

$$\mathrm{Regret}(\widehat{\pi}) = \mathcal{J}(\theta^*, \widehat{\pi}) - \mathcal{J}(\theta^*, \pi^*) \leq \mathcal{J}(\theta^*, \pi^*) - \mathcal{J}(\widehat{\theta}^*, \pi^*).$$

By Lemma 5.3 and the tower property of expectations, we have

$$\begin{aligned} \mathrm{Regret}(\widehat{\pi}) &\leq \sum_{h\in[H]} \mathbb{E}_{\theta^*,\pi^*}\big[(\Delta\mathbb{B}^{(\theta^*,\widehat{\theta}),\pi^*}_{h} V^{\widehat{\theta},\pi^*}_{h+1})(\tau_h)\big] \\ &= \mathbb{E}_{\theta^*,\pi^*}\Big[\sum_{h\in[H]} \mathbb{E}_{\theta^*,\pi^*}\big[(\Delta\mathbb{B}^{(\theta^*,\widehat{\theta}),\pi^*}_{h} V^{\widehat{\theta},\pi^*}_{h+1})(\tau_h) \,\big|\, s_{h-1}\big]\Big] = \mathbb{E}_{\theta^*,\pi^*}\Big[\sum_{h\in[H]} \epsilon_h(s_{h-1})\Big], \end{aligned}$$

where ϵ is defined in (5.8). Next, we invoke Assumption 4.4 to change the measure from $\mu^{\pi^*}_{h}$ to $\mu^{\overline{\pi}}_{h}$ for all $h \in [H]$. We have

$$\begin{aligned} \mathrm{Regret}(\widehat{\pi}) &= \int_{\mathcal{S}} \sum_{h\in[H]} \epsilon_h(s_{h-1}) \cdot \mu^{\pi^*}_{h}(s_{h-1})\, \mathrm{d}s_{h-1} = \int_{\mathcal{S}} \sum_{h\in[H]} \epsilon_h(s_{h-1}) \cdot \frac{\mu^{*}_{h}(s_{h-1})}{\mu^{\overline{\pi}}_{h}(s_{h-1})} \cdot \mu^{\overline{\pi}}_{h}(s_{h-1})\, \mathrm{d}s_{h-1} \\ &\leq \overline{C} \cdot \sum_{h\in[H]} \mathbb{E}_{\theta^*,\overline{\pi}}\big[\epsilon_h(s_{h-1})\big] \leq 2\overline{C} H^2 |\mathcal{A}^2| \gamma^2 \beta \cdot N^{-1/2}, \end{aligned}$$

where the last inequality follows from Lemma 5.4. □

C. Proofs for Lemmas in §5

C.1. Distribution Estimation Error

Proof of Lemma 5.1. For any $h \in [H]$ and $a, a' \in \mathcal{A}^2$, we have

$$
\begin{aligned}
\|\widehat{P}^{\dagger}_{h,a,a'} - P^{\dagger}_{h,a,a'}\|_1 &= \int_{\mathcal{O}^3} |\widetilde{\varphi}(o,o',o'')^\top \widehat{w}_{h,a,a'} - \widetilde{\varphi}(o,o',o'')^\top w_{h,a,a'}| \\
&\le \|\widehat{w}_{h,a,a'} - w_{h,a,a'}\|_2 \cdot \int_{\mathcal{O}^3} \|\widetilde{\varphi}(o,o',o'')\|_2 \,\mathrm{d}o\,\mathrm{d}o'\,\mathrm{d}o'' \\
&\le \sqrt{d} \cdot \|\widehat{w}_{h,a,a'} - w_{h,a,a'}\|_2. \qquad \text{(C.1)}
\end{aligned}
$$

Here, the first equality follows from Lemma 3.2, the first inequality follows from the Hölder's inequality, and the last inequality follows from Lemma D.1(a). Since we have closed-form of $\widehat{w}_{h,a,a'}$ in (3.13), we write

$$
\|\widehat{w}_{h,a,a'} - w_{h,a,a'}\|_2 = \|\Upsilon^{-1}U - \Upsilon^{-1}\Upsilon w_{h,a,a'}\|_2 \le \alpha^{-1} \cdot \|U - \Upsilon w_{h,a,a'}\|_2, \qquad \text{(C.2)}
$$

where Υ and U are defined in (3.11) and (3.12), respectively, and the last inequality follows from Assumption 4.5. We further examine $\Upsilon w_{h,a,a'}$ as follows

$$
[\Upsilon w_{h,a,a'}]_i = \sum_{j\in[d]} \big\langle \mathbb{K}\widetilde{\varphi}_i, \mathbb{K}\widetilde{\varphi}_j \big\rangle_{\mathcal{H}} \cdot [w_{h,a,a'}]_j = \Big\langle \mathbb{K}\widetilde{\varphi}_i, \mathbb{K}\Big(\sum_{j\in[d]} \phi_j \cdot [w_{h,a,a'}]_j\Big)\Big\rangle_{\mathcal{H}} = \langle \mathbb{K}\widetilde{\varphi}_i, \mathbb{K}P^{\dagger}_{h,a,a'}\rangle_{\mathcal{H}},
$$

where the second equality follows from the linearity of $\mathbb{K}$ and the last equality follows from the definition of $P^{\dagger}_{h,a,a'}$ in (3.9). Together with the definition of U in (3.12), we have

$$
\begin{aligned}
\|U - \Upsilon w_{h,a,a'}\|_2^2 &= \sum_{i\in[d]} \langle \mathbb{K}\widetilde{\varphi}_i, \widehat{\mathbb{K}}\mathcal{D}_{h,a,a'} - \mathbb{K}P^{\dagger}_{h,a,a'}\rangle_{\mathcal{H}}^2 \\
&\le \sum_{i\in[d]} \|\mathbb{K}\widetilde{\varphi}_i\|_{\mathcal{H}}^2 \cdot \|\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'} - \mathbb{K}P^{\dagger}_{h,a,a'}\|_{\mathcal{H}}^2 \le d \cdot \|\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'} - \mathbb{K}P^{\dagger}_{h,a,a'}\|_{\mathcal{H}}^2, \qquad \text{(C.3)}
\end{aligned}
$$

where the first inequality follows from the Cauchy-Schwarz inequality and the second inequality follows from Lemma D.1(b). Note that we can write

$$
\begin{aligned}
P^{\dagger}_{h,a_{h-1},a_h}(o_{h-1}, o_h, o_{h+1}) &= \mathbb{P}_{\theta,\pi^{\dagger}_{h,a_{h-1},a_h}}(o_{h-1}, a_{h-1}, o_h, a_h, o_{h+1}) \\
&= \mathbb{E}_{\theta,\pi^{\dagger}_{h,a_{h-1},a_h}}\Big[\mathbb{P}_{\theta,\pi^{\dagger}_{h,a_{h-1},a_h}}(o_{h-1}, a_{h-1}, o_h, a_h, o_{h+1} \,|\, \tau_{h-2}, a_{h-2})\Big],
\end{aligned}
$$

where $\pi^{\dagger}_{h,a_{h-1},a_h}$ is the same as the behavior policy except that it only chooses a_{h-1} at step $h-1$ and a_h at step h regarding the history. Since

$$
\begin{aligned}
&\mathbb{P}_{\theta,\bar{\pi}}(o_{h-1}, a_{h-1}, o_h, a_h, o_{h+1} \,|\, \tau_{h-2}, a_{h-2}) \\
&\quad = \mathbb{P}_{\theta,\pi^{\dagger}_{h,a_{h-1},a_h}}(o_{h-1}, a_{h-1}, o_h, a_h, o_{h+1} \,|\, \tau_{h-2}, a_{h-2}) \\
&\qquad \cdot \bar{\pi}(a_{h-1} \,|\, \tau_{h-2}, a_{h-2}, o_{h-1}) \cdot \bar{\pi}(a_h \,|\, \tau_{h-2}, a_{h-2}, o_{h-1}, a_{h-1}, o_h),
\end{aligned}
$$

we have

$$
P^{\dagger}_{h,a_{h-1},a_h}(o_{h-1}, o_h, o_{h+1}) = \mathbb{E}_{\theta,\bar{\pi}}\bigg[\frac{\mathbb{P}_{\theta,\bar{\pi}}(o_{h-1}, a_{h-1}, o_h, a_h, o_{h+1} \,|\, \tau_{h-2}, a_{h-2})}{\bar{\pi}(a_{h-1} \,|\, \tau_{h-2}, a_{h-2}, o_{h-1}) \cdot \bar{\pi}(a_h \,|\, \tau_{h-2}, a_{h-2}, o_{h-1}, a_{h-1}, o_h)}\bigg]
$$

Thus, a trivial unbiased estimator of $P^{\dagger}_{h,a,a'}$ would be

$$
\begin{aligned}
p^{\dagger}_{h,a,a'}(o,o',o'') &:= \sum_{(o_{h-1},o_h,o_{h+1},a_{h-1},a_h,\rho)\in\bigcup_{a,a'}\mathcal{D}_{h,a,a'}} \frac{\mathbb{1}\{(o_{h-1}, o_h, o_{h+1}, a_{h-1}, a_h) = (o, o', o'', a, a')\}}{N\rho} \qquad \text{(C.4)} \\
&= \sum_{(o_{h-1},o_h,o_{h+1},\rho)\in\mathcal{D}_{h,a,a'}} \frac{\mathbb{1}\{(o_{h-1}, o_h, o_{h+1}) = (o, o', o'')\}}{N\rho}. \qquad \text{(C.5)}
\end{aligned}
$$

We can rewrite $\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'}$ with the estimator $p^{\dagger}_{h,a,a'}$.

$$\begin{aligned}\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'}(x) &= \sum_{(x',\rho)\in\mathcal{D}_{h,a,a'}} \frac{\mathcal{K}(x',x)}{N\rho} = \int_{\mathcal{O}^3} \sum_{(x',\rho)\in\mathcal{D}_{h,a,a'}} \frac{\mathcal{K}(x',x)}{N\rho}\cdot \mathbb{1}\{x'=x''\}\,\mathrm{d}x'' \\ &= \int_{\mathcal{O}^3} \mathcal{K}(x'',x)\cdot p^{\dagger}_{h,a,a'}(x'')\,\mathrm{d}x'' = (\mathbb{K}p^{\dagger}_{h,a,a'})(x).\end{aligned}$$

Then, $\widehat{P}^{\dagger}_{h,a,a'}$ is an unbiased estimator of $P^{\dagger}_{h,a,a'}$ in that

$$\begin{aligned}\mathbb{E}\big[(\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'})(x)\big] &= \mathbb{E}\Big[\int_{\mathcal{O}^3} \mathcal{K}(x'',x)\cdot p^{\dagger}_{h,a,a'}(x'')\,\mathrm{d}x''\Big] = \int_{\mathcal{O}^3} \mathcal{K}(x'',x)\cdot \mathbb{E}\big[p^{\dagger}_{h,a,a'}(x'')\big]\,\mathrm{d}x'' \\ &= \int_{\mathcal{O}^3} \mathcal{K}(x'',x)\cdot P^{\dagger}_{h,a,a'}(x'')\,\mathrm{d}x'' = (\mathbb{K}P^{\dagger}_{h,a,a'})(x).\end{aligned}$$

In what follows, we upper bound $\|\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'} - \mathbb{K}P^{\dagger}_{h,a,a'}\|^2_{\mathcal{H}}$. We write

$$(\widehat{\mathbb{K}}_{h,a,a'}\{o^n_{h-1}, o^n_h, o^n_{h+1}, a^n_{h-1}, a^n_h, \rho^n\})(o,o',o'') = \mathcal{K}\big((o^n_{h-1}, o^n_h, o^n_{h+1}), (o,o',o'')\big)\cdot \mathbb{1}\{(a^n_{h-1}, a^n_h) = (a,a')\}\cdot \rho^{-1}.$$

Note that

$$\begin{aligned}\mathbb{E}\big[\|\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'} - \mathbb{K}P^{\dagger}_{h,a,a'}\|^2_{\mathcal{H}}\big] &\le \sqrt{\mathbb{E}\Big[\Big\|\frac{1}{N}\sum_{n\in[N]} \widehat{\mathbb{K}}_{h,a,a'}\{o^n_{h-1}, o^n_h, o^n_{h+1}, a^n_{h-1}, a^n_h, \rho^n_h\} - \mathbb{K}P^{\dagger}_{h,a,a'}\Big\|^2_{\mathcal{H}}\Big]} \\ &= \frac{1}{N}\sqrt{\mathbb{E}\Big[\Big\|\sum_{n\in[N]} \widehat{\mathbb{K}}\{o^n_{h-1}, o^n_h, o^n_{h+1}, a^n_{h-1}, a^n_h, \rho^n_h\}\Big\|^2_{\mathcal{H}}\Big] - \mathbb{E}\Big[\Big\|\sum_{n\in[N]} \mathbb{K}P^{\dagger}_{h,a,a'}\Big\|^2_{\mathcal{H}}\Big]},\end{aligned}$$

where the equality follows from

$$\mathbb{E}\Big[\Big\|\sum_{n\in[N]} \widehat{\mathbb{K}}\{o^n_{h-1}, o^n_h, o^n_{h+1}, a^n_{h-1}, a^n_h, \rho^n\}\Big\|_{\mathcal{H}}\Big] = \Big\|\sum_{n\in[N]} \mathbb{K}P^{\dagger}_{h,a,a'}\Big\|_{\mathcal{H}}.$$

Let $\bar{\rho}^n_{h,a,a'} = \mathbb{1}\{(a^n_{h-1}, a^n_h) = (a,a')\}\cdot(\rho^n_h)^{-1}$.

$$\begin{aligned}&\mathbb{E}\big[\|\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'} - \mathbb{K}P^{\dagger}_{h,a,a'}\|^2_{\mathcal{H}}\big] \\ &\quad\le \frac{1}{N}\sqrt{\sum_{n,n'\in[N]} \mathbb{E}\Big[\mathcal{K}\big((o^n_{h-1}, o^n_h, o^n_{h+1}), (o^{n'}_{h-1}, o^{n'}_h, o^{n'}_{h+1})\big)\cdot\bar{\rho}^n_{h,a,a'}\cdot\bar{\rho}^{n'}_{h,a,a'}\Big] - \mathbb{E}_{x,x'\overset{\text{i.i.d.}}{\sim} P^{\dagger}_{h,a,a'}}\big[\mathcal{K}(x,x')\big]} \\ &\quad= \frac{1}{N}\sqrt{\sum_{n\in[N]} \mathbb{E}\big[\mathcal{K}\big((o^n_{h-1}, o^n_h, o^n_{h+1}), (o^n_{h-1}, o^n_h, o^n_{h+1})\big)\cdot\bar{\rho}^n_{h,a,a'}\cdot\bar{\rho}^n_{h,a,a'}\big] - \mathbb{E}_{x\sim P^{\dagger}_{h,a,a'}}\big[\mathcal{K}(x,x)\big]}.\end{aligned}$$

Since $\mathcal{K}(x,x')\le 1$ for any $x,x'\in\mathcal{O}^3$ and $\pi(\cdot\,|\,\cdot)\ge \underline{C}$ by Assumption 4.4, we have

$$\mathbb{E}\big[\|\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'} - \mathbb{K}P^{\dagger}_{h,a,a'}\|^2_{\mathcal{H}}\big] \le \sqrt{2/N}/\underline{C}.$$

Finally, we invoke McDiarmid's inequality with $c_i = 2/(N\underline{C}^2)$. It holds with probability at least $1-\delta$ that

$$\big|\|\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'} - \mathbb{K}P^{\dagger}_{h,a,a'}\|^2_{\mathcal{H}} - \mathbb{E}\big[\|\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'} - \mathbb{K}P^{\dagger}_{h,a,a'}\|^2_{\mathcal{H}}\big]\big| \le \sqrt{2\log(2/\delta)/N}/\underline{C}.$$

Then, by the triangle inequality, it holds with probability at least $1-\delta$ that

$$\begin{aligned}\|\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'} - \mathbb{K}P^{\dagger}_{h,a,a'}\|^2_{\mathcal{H}} &\le \mathbb{E}\big[\|\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'} - \mathbb{K}P^{\dagger}_{h,a,a'}\|^2_{\mathcal{H}}\big] + \big|\|\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'} - \mathbb{K}P^{\dagger}_{h,a,a'}\|^2_{\mathcal{H}} - \mathbb{E}\big[\|\widehat{\mathbb{K}}\mathcal{D}_{h,a,a'} - \mathbb{K}P^{\dagger}_{h,a,a'}\|^2_{\mathcal{H}}\big]\big| \\ &\le \sqrt{10\log(2/\delta)/N}/\underline{C}. \qquad \text{(C.6)}\end{aligned}$$

By Combining (C.1), (C.2), (C.3) and (C.6), we have

$$\|\widehat{P}^{\ddagger}_{h,a} - P^{\ddagger}_{h,a}\|_1 \le \alpha^{-1} \cdot \sqrt{10d \cdot \log(2/\delta)/\underline{C}} \cdot N^{-1/2}.$$

By taking a union bound for all $(h, a, a') \in [H] \times \mathcal{A}^2$, we have

$$\|\widehat{P}^{\ddagger}_{h,a} - P^{\ddagger}_{h,a}\|_1 \le \alpha^{-1} \cdot \sqrt{10d \cdot \log(2H|\mathcal{A}|^2/\delta)/\underline{C}} \cdot N^{-1/2}$$

for all $h \in [H]$ and $a, a' \in \mathcal{A}$ with probability at least $1 - \delta$, which concludes the proof of (5.1). For (5.2), we note that

$$\begin{aligned}\|\widehat{P}^{\ddagger}_{h,a} - P^{\ddagger}_{h,a}\|_1 &= \int_{\mathcal{O}^2} \Big| \int_{\mathcal{O}} \big(\phi(o, o', o'')^\top \widehat{w}_{h,a,a'} - \phi(o, o', o'')^\top w\big) \, \mathrm{d}o'' \Big| \, \mathrm{d}o \, \mathrm{d}o' \\ &\le \int_{\mathcal{O}^3} |\phi(o, o', o'')^\top \widehat{w}_{h,a,a'} - \phi(o, o', o'')^\top w_{h,a,a}| \, \mathrm{d}o'' \, \mathrm{d}o \, \mathrm{d}o' = \|\widehat{P}^{\dagger}_{h,a} - P^{\dagger}_{h,a}\|_1,\end{aligned}$$

which holds when (5.1) holds. □

C.2. Confidence Interval

Proof of Lemma 5.2. We condition our proof on the event E defined in Lemma 5.1. By (**??**) and the triangle inequality, we have

$$\begin{aligned}\|\mathbb{F}^{\theta^*}_{h,a'} \widehat{P}^{\ddagger}_{h,a} - \widehat{P}^{\dagger}_{h,a,a'}\|_1 &= \|\mathbb{F}^{\theta^*}_{h,a'} \widehat{P}^{\ddagger}_{h,a} - \mathbb{F}^{\theta^*}_{h,a'} P^{\ddagger}_{h,a} + P^{\dagger}_{h,a,a'} - \widehat{P}^{\dagger}_{h,a,a'}\|_1 \\ &\le \|\mathbb{F}^{\theta^*}_{h,a'} (\widehat{P}^{\ddagger}_{h,a} - P^{\ddagger}_{h,a})\|_1 + \|P^{\dagger}_{h,a,a'} - \widehat{P}^{\dagger}_{h,a,a'}\|_1.\end{aligned} \tag{C.7}$$

By Lemma D.1(c), we have

$$\|\mathbb{F}^{\theta^*}_{h,a'} (\widehat{P}^{\ddagger}_{h,a} - P^{\ddagger}_{h,a})\|_1 \le \gamma \|\widehat{P}^{\ddagger}_{h,a,a'} - P^{\ddagger}_{h,a,a'}\|_1. \tag{C.8}$$

Plugging (C.7) into (C.8), it holds for all $(h, a, a') \in [H] \times \mathcal{A}^2$ that

$$\|\mathbb{F}^{\theta^*}_{h,a} \widehat{P}^{\ddagger}_{h,a} - \widehat{P}^{\dagger}_{h,a,a'}\|_1 \le (\gamma + 1) \cdot \|\widehat{P}^{\ddagger}_{h,a,a'} - P^{\ddagger}_{h,a,a'}\|_1 \le \beta \cdot N^{-1/2},$$

where the last inequality follows from Lemma 5.1 and the value of β is given in Theorem 4.6. Recall the definition of the confidence region in (**??**), we obtain $\theta^* \in \widehat{\Theta}$, which concludes the proof. □

C.3. Proof of Lemma 5.3

Proof. Let $\theta, \theta' \in \Theta$ be arbitrary. By the definition of R in (4.3), we have

$$\mathcal{J}(\theta, \pi) = \mathbb{E}_{\theta,\pi}\big[\sum_{h \in [H]} r(s_h, a_h)\big] = \mathbb{E}_{\theta,\pi}\big[R(\tau_{H+1})\big].$$

Meanwhile, by Lemma 4.3 and the fact that $\mathbb{P}_\theta(\tau_1) = \mathbb{P}_{\theta'}(\tau_1)$ for any $\tau_1 \in \Gamma_1$, we have

$$\mathcal{J}(\theta', \pi) = \mathbb{E}_{\theta',\pi}\big[\sum_{h \in [H]} r(s_h, a_h)\big] = \mathbb{E}_{\theta'}\big[V_1^{\theta',\pi}(\tau_1)\big] = \mathbb{E}_{\theta}\big[V_1^{\theta',\pi}(\tau_1)\big]$$

Then, by the additional definition that $V^{\theta,\pi}_{H+1} = V^{\theta',\pi}_{H+1} = R$, we have

$$\mathcal{J}(\theta, \pi) - \mathcal{J}(\theta', \pi) = \mathbb{E}_{\theta,\pi}\big[V^{\theta',\pi}_{H+1}(\tau_{H+1}) - V^{\theta',\pi}_1(\tau_1)\big] = \sum_{h \in [H]} \mathbb{E}_{\theta,\pi}\big[V^{\theta',\pi}_{h+1}(\tau_{h+1}) - V^{\theta',\pi}_h(\tau_h)\big] \tag{C.9}$$

By Lemma 4.1, we have

$$\mathbb{E}_{\theta,\pi}\big[V^{\theta',\pi}_{h+1}(\tau_{h+1})\big] = \mathbb{E}_{\theta,\pi}\big[(\mathbb{B}^{\theta,\pi}_h V^{\theta',\pi}_{h+1})(\tau_h)\big]. \tag{C.10}$$

Meanwhile, by the definition of $V_h^{\theta,\pi}$ in (4.2), we have

$$\mathbb{E}_{\theta,\pi}\big[V_h^{\theta',\pi}(\tau_h)\big] = \mathbb{E}_{\theta,\pi}\big[(\mathbb{B}_h^{\theta',\pi} V_{h+1}^{\theta',\pi})(\tau_h)\big]. \tag{C.11}$$

Plugging (C.11) and (C.10) into (C.9), we have

$$\mathcal{J}(\theta,\pi) - \mathcal{J}(\theta',\pi) = \sum_{h\in[H]} \mathbb{E}_{\theta,\pi}\big[(\mathbb{B}_h^{\theta,\pi} V_{h+1}^{\theta',\pi})(\tau_h) - (\mathbb{B}_h^{\theta',\pi} V_{h+1}^{\theta',\pi})(\tau_h)\big],$$

which concludes the proof of Lemma 5.3. □

C.4. Proof of Lemma 5.4

Proof. For notational simplicity, we denote

$$\Delta\mathcal{F}_{h,a}(o,o',o'') = \mathcal{F}_{h,a}^{\theta^*}(o,o',o'') - \mathcal{F}_{h,a}^{\widehat{\theta}}(o,o',o''). \tag{C.12}$$

Then, we invoke Lemma 4.2 and rewrite the value function as follows

$$\begin{aligned}
\epsilon_h(s_{h-1}) &= \Big|\mathbb{E}_{\theta^*,\pi^*}\Big[\big((\mathbb{B}_h^{\theta^*,\pi^*} - \mathbb{B}_h^{\widehat{\theta},\pi^*})V_{h+1}^{\widehat{\theta},\pi^*}\big)(\tau_h^*)\Big]\Big| \\
&= \Big|\mathbb{E}_{\theta^*,\pi^*}\Big[\int_{\mathcal{O}^2\times\mathcal{S}} \sum_{a_h\in\mathcal{A}} \mathbb{E}_{\theta^*,\pi^*}\Big[\sum_{i\in[H]} r(o_i,a_i)\,\Big|\, s_{h+1},\tau_{h-1}^*,a_{h-1}^*,o_h,a_h\Big] \\
&\qquad \cdot \xi_h^{\widehat{\theta}}(s_{h+1},o_{h+1}) \cdot \Delta\mathcal{F}_{h,a_h}(o_h^*,o_h,o_{h+1}) \cdot \pi^*(a_h \,|\, \tau_{h-1}^*,a_{h-1}^*,o_h)\,\mathrm{d}o_h\,\mathrm{d}o_{h+1}\,\mathrm{d}s_{h+1}\Big]\Big|,
\end{aligned}$$

where $\Delta\mathcal{F}_{h,a_h}$ is defined in (C.12). To handle the tricky expectation taken with respect to τ_h, we take advantage of the boundedness of r and π and the fact that o_h^* only depend on s_h. By Jensen's inequality, we have

$$\begin{aligned}
\epsilon_h(s_{h-1}) &\le H \cdot \sum_{a_{h-1},a_h\in\mathcal{A}} \int_{\mathcal{O}\times\mathcal{S}} \Big|\int_{\mathcal{O}} \xi_h^{\widehat{\theta}}(s_{h+1},o_{h+1}) \cdot \mathbb{E}_{o_h^*\sim\mathbb{P}_{\theta^*}(\cdot\,|\,s_{h-1},a_{h-1})}\big[\Delta\mathcal{F}_{h,a_h}(o_h^*,o_h,o_{h+1})\big]\,\mathrm{d}o_{h+1}\Big|\,\mathrm{d}o_h\,\mathrm{d}s_{h+1} \\
&\le H\cdot\gamma \sum_{a_{h-1},a_h\in\mathcal{A}} \int_{\mathcal{O}^2} \big|\mathbb{E}_{o_h^*\sim\mathbb{P}_{\theta^*}(\cdot\,|\,s_{h-1},a_{h-1})}\big[\Delta\mathcal{F}_{h,a_h}(o_h^*,o_h,o_{h+1})\big]\big|\,\mathrm{d}o_h\,\mathrm{d}o_{h+1},
\end{aligned} \tag{C.13}$$

where the last inequality follows from Assumption 2.2 and the action a_{h-1} in the condition of the expectation is decoupled from history and the policy. The right-hand side of (C.13) does involve the optimal policy π^* at all. Since our target is $\mathbb{E}_{\theta^*,\bar{\pi}}[\epsilon_h(s_{h-1})]$, we have

$$\begin{aligned}
&\mathbb{E}_{\theta^*,\bar{\pi}}\Big[\big|\mathbb{E}_{o_h^*\sim\mathbb{P}_{\theta^*}(\cdot\,|\,s_{h-1},a_{h-1})}\big[\Delta\mathcal{F}_{h,a_h}(o_h^*,o_h,o_{h+1})\big]\big|\Big] \\
&\qquad = \int_{\mathcal{S}}\Big|\int_{\mathcal{O}} \Delta\mathcal{F}_{h,a_h}(o_h^*,o_h,o_{h+1}) \cdot \mathbb{P}_{\theta^*,\bar{\pi}}(s_{h-1},o_h^*\,|\,a_{h-1})\,\mathrm{d}o_h^*\Big|\,\mathrm{d}s_{h-1}.
\end{aligned} \tag{C.14}$$

Prior to deriving the upper bound of (C.14), we note that for any $(o_h,o_{h+1},a_h)\in\mathcal{O}^2\times\mathcal{A}$,

$$\begin{aligned}
b_{h,o_h,o_h^*,o_{h+1},a_{h-1},a_h}(s_{h-1}) &:= \int_{\mathcal{O}} \Delta\mathcal{F}_{h,a_h}(o_h^*,o_h,o_{h+1}) \cdot \mathbb{P}_{\theta^*,\bar{\pi}}(s_{h-1},o_h^*\,|\,a_{h-1})\,\mathrm{d}o_h^* \\
&= \int_{\mathcal{O}} \Delta\mathcal{F}_{h,a_h}(o_h^*,o_h,o_{h+1}) \cdot \mathcal{I}_{h-1}^{\theta^*}(s_{h-1}\,|\,o_h^*,a_{h-1}) \cdot \mathbb{P}_{\theta^*,\bar{\pi}}(o_h^*\,|\,a_{h-1})\,\mathrm{d}o_h^*,
\end{aligned} \tag{C.15}$$

where $\mathcal{I}$ is the inverse transition function defined in §A.2.2, where we also show $\mathcal{I}_{h-1}^{\theta^*}(\cdot\,|\,o_h^*,a_{h-1})\in\mathrm{linspan}(\psi)$. Then, we have $b_{h,o_h,o_h^*,o_{h+1},a_{h-1},a_h}(\cdot)\in\mathrm{linspan}(\psi)$. Thus, by invoking Assumption 2.2, it holds that

$$\begin{aligned}
&\mathbb{E}_{\theta^*,\bar{\pi}}\Big[\big|\mathbb{E}_{o_h^*\sim\mathbb{P}_{\theta^*}(\cdot\,|\,s_{h-1},a_{h-1})}\big[\Delta\mathcal{F}_{h,a_h}(o_h^*,o_h,o_{h+1})\big]\big|\Big] \\
&\quad = \|b_{h,o_h,o_h^*,o_{h+1},a_{h-1},a_h}\|_1 = \|\mathbb{U}\mathbb{O}b_{h,o_h,o_h^*,o_{h+1},a_{h-1},a_h}\|_1 \le \gamma\cdot\|\mathbb{O}b_{h,o_h,o_h^*,o_{h+1},a_{h-1},a_h}\|_1.
\end{aligned}$$

By the definition of $b_{h,o_h,o_h^*,o_{h+1},a_{h-1},a_h}(s_{h-1})$ in (C.15) and the definition of the operator $\mathbb{F}$ in (3.2), we have

$$\begin{aligned}
&\mathbb{E}_{\theta^*,\bar{\pi}}\Big[\Big|\mathbb{E}_{o_h^*\sim\mathbb{P}_{\theta^*}(\cdot\,|\,s_{h-1},a_{h-1})}\big[\Delta\mathcal{F}_{h,a_h}(o_h^*,o_h,o_{h+1})\big]\Big|\Big] \\
&\quad\le \gamma\cdot\int_{\mathcal{O}}\Big|\int_{\mathcal{O}}\Delta\mathcal{F}_{h,a_h}(o_h^*,o_h,o_{h+1})\cdot\mathbb{P}_{\theta^*,\bar{\pi}}(o_{h-1},o_h^*\,|\,a_{h-1})\,\mathrm{d}o_h^*\Big|\,\mathrm{d}o_{h-1} \\
&\quad= \gamma\cdot\int_{\mathcal{O}}|(\mathbb{F}^{\widehat{\theta}}_{h,a_h}P^{\ddagger}_{h,a_{h-1}})(o_{h-1},o_h,o_{h+1})-(\mathbb{F}^{\theta^*}_{h,a_h}P^{\ddagger}_{h,a_{h-1}})(o_{h-1},o_h,o_{h+1})|\,\mathrm{d}o_{h-1}. \qquad\text{(C.16)}
\end{aligned}$$

Combining (C.13) and (C.16), we have

$$\mathbb{E}_{\theta^*,\bar{\pi}}[\epsilon_h]\le H\gamma^2\cdot\sum_{a,a'\in\mathcal{A}}\|\mathbb{F}^{\widehat{\theta}}_{h,a'}P^{\ddagger}_{h,a}-P^{\dagger}_{h,a,a'}\|_1. \qquad\text{(C.17)}$$

Next, by the triangle inequality, we have

$$\begin{aligned}
&\|\mathbb{F}^{\widehat{\theta}}_{h,a'}P^{\ddagger}_{h,a}-P^{\dagger}_{h,a,a'}\|_1 \\
&\quad\le \underbrace{\|\mathbb{F}^{\widehat{\theta}}_{h,a'}P^{\ddagger}_{h,a}-\mathbb{F}^{\widehat{\theta}}_{h,a'}\widehat{P}^{\ddagger}_{h,a}\|_1}_{\text{(i)}}+\underbrace{\|\mathbb{F}^{\widehat{\theta}}_{h,a'}\widehat{P}^{\ddagger}_{h,a}-\widehat{P}^{\dagger}_{h,a,a'}\|_1}_{\text{(ii)}}+\underbrace{\|\widehat{P}^{\dagger}_{h,a,a'}-P^{\dagger}_{h,a,a'}\|_1}_{\text{(iii)}}. \qquad\text{(C.18)}
\end{aligned}$$

By Lemma D.1(c), Lemma 5.1, and the definition of the confidence region in (**??**), we have

$$\text{(i)}\le\gamma\cdot\|P^{\ddagger}_{h,a}-\widehat{P}^{\ddagger}_{h,a}\|_1\le\gamma\cdot\beta_0\cdot N^{-1/2}, \qquad\text{(C.19)}$$

$$\text{(ii)}\le\beta\cdot N^{-1/2}, \qquad\text{(C.20)}$$

$$\text{(iii)}\le\beta_0\cdot N^{-1/2}. \qquad\text{(C.21)}$$

Combining (C.17) - (C.21), we have

$$\mathbb{E}_{\theta^*,\bar{\pi}}[\epsilon_h]\le 2H|\mathcal{A}^2|\gamma^2\beta\cdot N^{-1/2},$$

which concludes the proof of Lemma 5.4. □

D. Auxiliary Lemmas

D.1. Upper Bounds

Lemma D.1. *We have the following inequalities which are used in our proofs and analysis.*

(a). It holds for any $o,o',o''\in\mathcal{O}$ that $\|\widetilde{\varphi}(o,o',o'')\|_2\le\sqrt{d}$.

(b). Under Assumption 4.5, it holds for any function $p:\mathcal{O}^3\to[0,1]$ that $\|\mathbb{K}p\|^2_{\mathcal{H}}\le 1$.

(c). It holds for any $\ell\in L^1(\mathcal{O}^2)$, $h\in[H]$ and $\theta\in\Theta$ that $\|\mathbb{F}^{\theta}_{h,a}f\|_1\le\gamma\|f\|_1$.

Proof. To prove (a), following from the fact that entries of ϕ are probability distributions, we have

$$\|\widetilde{\varphi}(o,o',o'')\|_2^2=\sum_{(i,j,k)\in[d_\varphi]^3}\phi_i(o)^2\phi_j(o')^2\phi_k(o'')^2\le d.$$

To prove (b), we invoke Assumption 4.5, which tells us $|\mathcal{K}(\cdot,\cdot)|\le 1$. Meanwhile, we have $|p(\cdot)|\le 1$ by assumption. Thus, by definition, we naturally have

$$\|\mathbb{K}p\|^2_{\mathcal{H}}=\int_{\mathcal{O}^3\times\mathcal{O}^3}\mathcal{K}(x,y)p(x)p(y)\,\mathrm{d}x\,\mathrm{d}y\le 1.$$

To prove (c), we have

$$\begin{aligned}
\|\mathbb{F}^{\theta}_{h,a_h} f\|_1 &= \int_{\mathcal{O}^3} \Big| \int_{\mathcal{O}} f(o_{h-1}, o_h') \cdot \mathcal{F}^{\theta}_{h,a_h}(o_h', o_h, o_{h+1}) \, \mathrm{d}o_h' \Big| \, \mathrm{d}o_{h-1} \, \mathrm{d}o_h \, \mathrm{d}o_{h+1} \\
&= \int_{\mathcal{O}^3} \Big| \int_{\mathcal{S}\times\mathcal{O}} f(o_{h-1}, o_h') \cdot \mathbb{P}_{\theta}(o_h, o_{h+1} \,|\, s_h, a_h) \cdot \xi^{\theta}_h(s_h, o_h') \, \mathrm{d}s_h \, \mathrm{d}o_h' \Big| \, \mathrm{d}o_{h-1} \, \mathrm{d}o_h \, \mathrm{d}o_{h+1} \\
&\le \int_{\mathcal{O}^3\times\mathcal{S}} \Big| \int_{\mathcal{O}} f(o_{h-1}, o_h') \cdot \xi^{\theta}_h(s_h, o_h') \, \mathrm{d}o_h' \Big| \cdot \mathbb{P}_{\theta}(o_h, o_{h+1} \,|\, s_h, a_h) \, \mathrm{d}o_{h-1} \, \mathrm{d}o_h \, \mathrm{d}o_{h+1} \, \mathrm{d}s_h \\
&\le \int_{\mathcal{O}^3\times\mathcal{S}} \Big| \int_{\mathcal{O}} f(o_{h-1}, o_h') \cdot \xi^{\theta}_h(s_h, o_h) \, \mathrm{d}o_h \Big| \, \mathrm{d}o_{h-1} \, \mathrm{d}s_h \\
&= \gamma \cdot \|f\|_1,
\end{aligned}$$

where the first inequality follows from the Jensen's inequality, and the second inequality follows from Assumption 2.2. □

D.2. Concentration Inequalities

Lemma D.2. *Let $\mu_1, \ldots, \mu_t$ be T distributions over $\mathcal{X}$ and $\mathcal{K}$ be a kernel function over $\mathcal{X} \times \mathcal{X}$ that satisfies $|\mathcal{K}(x, x')| \le 1$ for any $x, x' \in \mathcal{X}$. Suppose X_t is independently sampled from μ_t for any $t \in [T]$. Define function $y : \mathcal{X}^n \to \mathbb{R}$ to be*

$$y(X_1, \ldots, X_n) = \Big\| \frac{1}{T} \sum_{t\in[T]} \mathbb{K}\delta_{X_t} - \frac{1}{T} \sum_{t\in[T]} \mathbb{K}\mu_t \Big\|_{\mathcal{H}},$$

where δ is the Dirac delta function, $\mathbb{K}$ is the embedding operator of $\mathcal{K}$, and $\mathcal{H}$ is the RKHS induced by the kernel $\mathcal{K}$. Then, it holds for any $\delta > 0$ with probability at least $1 - \delta$ that

$$y(X_1, \ldots, X_n) \le (\log(1/\delta) + 1) \cdot T^{-1/2}.$$

Proof. By Jensen's inequality, we have

$$\begin{aligned}
&\mathbb{E}\big[y(X_1, \ldots, X_n)\big] \\
&\quad \le \sqrt{\mathbb{E}\Big[\Big\| \frac{1}{T} \sum_{t\in[T]} \mathbb{K}\delta_{X_t} - \frac{1}{T} \sum_{t\in[T]} \mathbb{K}\mu_t \Big\|^2_{\mathcal{H}}\Big]} \\
&\quad = \frac{1}{T} \sqrt{\mathbb{E}\Big[\Big\| \sum_{t\in[T]} \mathbb{K}\delta_{X_t} \Big\|^2_{\mathcal{H}}\Big] - \mathbb{E}\Big[\Big\| \sum_{t\in[T]} \mathbb{K}\delta_{X_t} \Big\|_{\mathcal{H}} \cdot \Big\| \sum_{t\in[T]} \mathbb{K}\mu_t \Big\|_{\mathcal{H}}\Big] + \mathbb{E}\Big[\Big\| \sum_{t\in[T]} \mathbb{K}\mu_t \Big\|^2_{\mathcal{H}}\Big]}.
\end{aligned}$$

By definition, it holds for any $t \in [T]$ that

$$\mathbb{E}\Big[\Big\| \sum_{t\in[T]} \mathbb{K}\delta_{X_t} \Big\|_{\mathcal{H}}\Big] = \Big\| \sum_{t\in[T]} \mathbb{K}\mu_t \Big\|_{\mathcal{H}}.$$

Thus, we have

$$\begin{aligned}
\mathbb{E}\big[y(X_1, \ldots, X_n)\big] &\le \frac{1}{T} \sqrt{\mathbb{E}\Big[\Big\| \sum_{t\in[T]} \mathbb{K}\delta_{X_t} \Big\|^2_{\mathcal{H}}\Big] - \Big\| \sum_{t\in[T]} \mathbb{K}\mu_t \Big\|^2_{\mathcal{H}}} \\
&\le \frac{1}{T} \sqrt{\sum_{t\in[T]} \sum_{t'\in[T]} \mathbb{E}[\mathcal{K}(X_t, X_{t'})] - \mathbb{E}_{X\sim\mu_t, X'\sim\mu_{t'}}[\mathcal{K}(X, X')]} \\
&= \frac{1}{T} \sqrt{\sum_{t\in[T]} \mathbb{E}[\mathcal{K}(X_t, X_t)] - \mathbb{E}_{X\sim\mu_t, X'\sim\mu_{t'}}[\mathcal{K}(X, X')]},
\end{aligned}$$

where the equality follows from the fact that X_t and $X_{t'}$ are independent for $t \neq t'$. Since $|\mathcal{K}(x, x')| \leq 1$ for any $x, x' \in \mathcal{X}$, we have

$$\mathbb{E}[y(X_1, \ldots, X_n)] \leq \sqrt{2/T}. \tag{D.1}$$

Also, for any $t' \in [T]$ and $x_1, \ldots, x_T, x'_{t'} \in \mathcal{X}$, we have

$$\begin{aligned}
&|y(x_1, \ldots, x_T) - y(x_1, \ldots, x_{t-1}, x_{t'}, x_{t'+1}, \ldots, x_T)| \\
&\quad = \frac{1}{T}\Big|\Big\| \sum_{t \in [T]} \mathbb{K}\delta_{x_t} - \mathbb{K}\mu_t \Big\|_{\mathcal{H}} - \Big\| \mathbb{K}\delta_{x'_{t'}} - \mathbb{K}\delta_{x_{t'}} + \sum_{t \in [T]} \mathbb{K}\delta_{x_t} - \mathbb{K}\mu_t \Big\|_{\mathcal{H}}\Big| \leq \frac{1}{T}\|\mathbb{K}\delta_{x_{t'}} - \mathbb{K}\delta_{x'_{t'}}\|_{\mathcal{H}},
\end{aligned}$$

where the inequality follows from the triangle inequality. By the fact that $|\mathcal{K}(x, x')| \leq 1$ for any $x, x' \in \mathcal{X}$, we have

$$\|\mathbb{K}\delta_{x_{t'}} - \mathbb{K}\delta_{x'_{t'}}\|_{\mathcal{H}}^2 = \mathcal{K}(x_{t'}, x_{t'}) - \mathcal{K}(x_{t'}, x'_{t'}) - \mathcal{K}(x'_{t'}, x_{t'}) + \mathcal{K}(x'_{t'}, x'_{t'}) \leq 4.$$

Then, by invoking the McDiarmid's inequality (see also Lemma D.3) with $c_i = 2/T$, we have with probability at least $1 - \delta$ that

$$\big|y(X_1, \ldots, X_n) - \mathbb{E}\big[y(X_1, \ldots, X_n)\big]\big| \leq \sqrt{2\log(2/\delta)/T}. \tag{D.2}$$

Then, by the triangle inequality and combining (D.1) and (D.2), it holds for any $\delta \in (0, 1)$ with probability at least $1 - \delta$ that

$$|y(X_1, \ldots, X_n)| \leq \big|\mathbb{E}\big[y(X_1, \ldots, X_n)\big]\big| + \big|y(X_1, \ldots, X_n) - \mathbb{E}\big[y(X_1, \ldots, X_n)\big]\big| \leq \sqrt{10\log(2/\delta)/T},$$

which concludes the proof of Lemma D.2. □

We put here the McDiarmid's inequality for reference without proving it.

Lemma D.3. *[McDiarmid's Inequality] Let $X_1, \ldots, X_n$ be independent random variables with ranges $\mathcal{X}_1, \ldots, \mathcal{X}_n$. Let $y : \mathcal{X}_1 \times \cdots \times \mathcal{X}_n \to \mathbb{R}$ be any function. If there exists constants $c_1, \ldots, c_n$ such that for any $i \in [n]$,*

$$|y(x_1, \ldots, x_n) - y(x'_1, \ldots, x'_n)| \leq c_i,$$

for any $(x_1, \ldots, x_n), (x'_1, \ldots, x'_n) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ that differ only in the i-th coordinate, then it holds for any $\epsilon > 0$ that

$$\mathbb{P}\Big(\big|y(X_1, \ldots, X_n) - \mathbb{E}\big[y(X_1, \ldots, X_n)\big]\big| \geq \epsilon\Big) \leq 2\exp\Big(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\Big).$$