Statistical Tractability of Off-policy Evaluation of History-dependent Policies in POMDPs

Anonymous authors
Paper under double-blind review

Abstract

We investigate off-policy evaluation (OPE), a central and fundamental problem in reinforcement learning (RL), in the challenging setting of Partially Observable Markov Decision Processes (POMDPs) with large observation spaces. Recent works of Uehara et al. (2023a); Zhang & Jiang (2024) developed a *model-free* framework and identified important coverage assumptions (called *belief* and *outcome* coverage) that enable accurate OPE of *memoryless* policies with polynomial sample complexities, but handling more general target policies that depend on the entire observable history remained an open problem. In this work, we prove information-theoretic hardness for model-free OPE of history-dependent policies in several settings, characterized by additional assumptions imposed on the behavior policy (memoryless vs. history-dependent) and/or the state-revealing property of the POMDP (single-step vs. multi-step revealing). We further show that some hardness can be circumvented by a natural *model-based* algorithm—whose analysis has surprisingly eluded the literature despite the algorithm's simplicity—demonstrating provable separation between model-free and model-based OPE in POMDPs.

1 Introduction

Off-policy evaluation (OPE) aims to evaluate a target policy π_e using an offline dataset collected by a different behavior policy π_b. The problem plays a crucial role in reinforcement learning (RL), and is particularly relevant to real-world scenarios where policies need to be properly evaluated before online deployment (Murphy, 2003; Ernst et al., 2006; Mandel et al., 2014; Bottou et al., 2013; Chapelle et al., 2014; Theocharous et al., 2015).

Efficient OPE requires the behavior policy π_b to satisfy certain coverage assumptions with respect to the target policy π_e. In the setting of Markov Decision Processes (MDPs), it is well established that a bounded state-action density ratio between π_e and π_b suffices for polynomial sample-complexity bounds; see Uehara et al. (2022b); Jiang & Xie (2024) for surveys and tutorials on the topic. However, the Markov assumption, that the immediate observation is a sufficient statistic of history, can be restrictive in scenarios where the state is latent and unobservable to the agent, as is often the case in many real-world applications.

In this paper, we study OPE in non-Markov environments modelled as partially observable MDPs (POMDPs),[1] where the observation space is large and demands the use of function approximation. In POMDPs, the agent only has access to observations rather than the latent state, and the next observation may depend on the entire history of observation-action sequences (or simply, the history). A common approach to apply MDP techniques is to treat the history as the state, thereby reducing a POMDP to a history-based MDP. However, under this conversion, the state-action density ratio becomes the density ratio of the entire observation-action sequence, which grows exponentially with the horizon length.

To address this issue, a recent line of research (Uehara et al., 2023a; Zhang & Jiang, 2024) has proposed model-free methods for OPE in POMDPs with large observation spaces. Zhang & Jiang

[1] When we refer to OPE in POMDPs, we mean *unconfounded* POMDPs, where the behavior policy only depends on the observable variables and not the latent state. There is also research on OPE in confounded POMDPs, where the behavior policy π_b depends on (*only*) the latent state (Shi et al., 2022; Bennett & Kallus, 2024); see Zhang & Jiang (2024) for further discussions on the distinction between the two settings.

Table 1: Summary of whether $\mathrm{poly}(H, \log(|\mathcal{M}|/\delta), \epsilon, C_{\mathcal{A}}, C_{\square}, C_{\mathcal{H}})$ (c.f. Theorem 1) complexity is achievable in different settings, where $C_{\square}$ is either $C_{\mathcal{O}}$ or $C_{\mathcal{F}}$ depending on whether single-step or multi-step revealing is assumed. "MF" and "MB" stand for model-free (Definition 2) and model-based (Section 4), respectively. "✓" indicates positive results, and "✗" indicates information-theoretic hardness. The setting is the easiest in the top-left corner, and becomes harder in the right or down direction. Therefore, the hardness of MF in Row 2 automatically implies those in Row 3. The bottom-right corner for MB is an open problem which we conjecture to be intractable.

Policy Types	**Single-step Revealing**	**Multi-step Revealing**
Memoryless π_b & π_e	MF: ✓ (Zhang & Jiang, 2024), MB: ✓ (Theorem 5)	
Memoryless π_b & History-dependent π_e	MF: ✗ (Theorem 3), MB: ✓ (Theorem 5)	
History-dependent π_b & π_e	MF: ✗ , MB: ✓ (Theorem 4)	MF: ✗ , MB: ?

(2024) identify two novel coverage assumptions for OPE in POMDPs, belief and outcome coverage, and demonstrate that their algorithm achieves polynomial sample complexity under these assumptions. However, they focus on evaluating *memoryless* target policies π_e, which ignore history and depend only on the current observation. Extension to history-dependent π_e exists, but the guarantees quickly deteriorate when the history window that π_e depends on has a nontrivial length (Uehara et al., 2023a, Appendix C). This motivates us to study the following question:

When can we achieve polynomial sample complexity for OPE of history-dependent target policies?

We investigate the question in a range of concrete settings, and the answer turns out to be more complex than a simple yes and no. These settings are defined by variations along several dimensions:

- **Model-free vs. model-based algorithms** The algorithms in Uehara et al. (2023a); Zhang & Jiang (2024) are *model-free*, in the sense that the algorithm only queries π_e on histories in the offline dataset. Under this very broad definition (Definition 3 in Zhang & Jiang (2024)), we show that ***no model-free algorithms can handle general history-dependent target policies***, even if we impose additional assumptions to make the problem easier in other dimensions (see below). This motivates us to also consider model-based algorithms that fit a POMDP model from data, which circumvent the hardness as they query π_e on model-generated synthetic trajectories.
- **Single-step vs. multi-step (outcome) revealing** The outcome coverage condition identified by Zhang & Jiang (2024) asserts that the future observation-action sequences can probabilistically decode the latent state (Assumption 9 in Zhang & Jiang (2024)). A stronger version of the condition is that the immediate observation suffices, which corresponds to a standard (single-step) "revealing" assumption commonly made in online RL for POMDPs (Liu et al., 2022a).
- **Memoryless vs. history-dependent π_b** Another dimension is whether π_b is also history dependent. In the literature, it has been reported that a history-dependent π_b often makes it difficult to infer POMDP dynamics from data, and a memoryless π_b makes it easier to do so (Kwon et al., 2024).

Our findings are summarized in Table 1. With either relaxation (single-step revealing *or* memoryless π_b), a simple Maximum Likelihood Estimation (MLE)-based model-based algorithm achieves desired guarantees, where all model-free algorithms must suffer hardness. To our best knowledge, these results are the first polynomial sample-complexity bound for evaluating history-dependent target policies under coverage assumptions, and demonstrate a formal separation between model-based and model-free OPE in POMDPs.

2 Preliminaries

Notation. For a vector $\mathbf{a}$, we use $\mathrm{diag}(\mathbf{a})$ to denote the diagonal matrix with $\mathbf{a}$ as the diagonal and use $[\mathbf{a}]_i$ to denote its i-th element. We use $\mathbf{e}_i$ to denote the basis vector with the i-th element being one. For a positive integer n, we use $[n]$ to denote the set $\{1, 2, \cdots, n\}$. For a matrix M, we use $\sigma_{\min}(M)$ and $M^\dagger$ to denote its minimum singular value and pseudo-inverse respectively. The ij entry of matrix M is denoted as $[M]_{ij}$ and the i-th row of M is denoted as $[M]_{i,:}$. The L_1 norm of matrix M is $\|M\|_1 = \sup_{\mathbf{x}\neq \mathbf{0}} \frac{\|M\mathbf{x}\|_1}{\|\mathbf{x}\|_1}$.

POMDP Setup. We use tuple $\left\langle H, \mathcal{S} = \bigcup_{h=1}^{H} \mathcal{S}_h, \mathcal{A}, \mathcal{O} = \bigcup_{h=1}^{H} \mathcal{O}_h, R, \mathbb{O}, \mathbb{T}, d_1 \right\rangle$ to specify a finite-horizon POMDP. Here H is the length of horizon; $\mathcal{S}_h$ is the state space at step h; $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$; $\mathcal{O}_h$ is the observation space at step h with $|\mathcal{O}_h| = O$; $R : \mathcal{O} \to [0, 1]$ is the reward function; $\mathbb{T} = \{\mathbb{T}_h\}_{h\in[H-1]}$ is the collection of transition dynamics where $\mathbb{T}_h : \mathcal{S}_h \times \mathcal{A} \to \mathcal{S}_{h+1}$; $\mathbb{O} = \{\mathbb{O}_h\}_{h\in[H]}$ is the collection of emission dynamics where $\mathbb{O}_h : \mathcal{S}_h \to \mathcal{O}_h$; $d_1 \in \Delta(\mathcal{S}_1)$ is the distribution of initial state s_1. All state, action, and observation spaces are finite and discrete. However, since the observation space can be very rich, the cardinality O may be arbitrarily large. Therefore, we aim to obtain sample complexity bounds that avoid any explicit dependence on O.

At the beginning of each episode in the POMDP, an initial state s_1 drawn from d_1. At each step h, the decision-making agent observes $o_h \sim \mathbb{O}_h(\cdot \mid s_h)$, along with the reward $r_h(o_h)$, and then takes an action a_h. After this, the environment transitions to the next state $s_{h+1} \sim \mathbb{T}_h(\cdot \mid s_h, a_h)$, and the episode terminates after a_H. Note that the states $s_{1:H}$ are latent and unobservable to the agent.

History and Belief State. We use $\tau_h = (o_1, a_1, \cdots, o_h, a_h) \in \mathcal{T}_h := \prod_{h'=1}^{h}(\mathcal{O}_h \times \mathcal{A})$ to denote the historical observation-action sequence up to step h. Given history τ_h, we define $\mathbf{b}_{\mathcal{S}}(\tau_h) \in \mathbb{R}^{|\mathcal{S}_{h+1}|}$ as its belief state vector and $[\mathbf{b}_{\mathcal{S}}(\tau_h)]_i = \mathbb{P}(s_{h+1} = i \mid \tau_h)$. [2]

Policies. A (history-dependent) policy $\pi = \{\pi_h\}_{h\in[H]}$ where $\pi_h : \mathcal{T}_{h-1} \times \mathcal{O}_h \to \Delta(\mathcal{A})$ specifies the action probability given the history τ_{h-1} and the current observation o_h. A *memoryless* policy only depends on the current observation (i.e., $\pi_h : \mathcal{O}_h \to \Delta(\mathcal{A})$). We define $J(\pi)$ as the expected cumulative return under policy π: $J(\pi) := \mathbb{E}_\pi \left[\sum_{h=1}^{H} R(o_h)\right]$. Here we use $\mathbb{E}_\pi$ to denote the expectations under policy π, $\mathbb{P}^\pi(\cdot)$ for the probability of an event under the same policy and $d_h^\pi(\cdot)$ for the marginal distribution of s_h, a_h under π.

The Outcome Matrix. For any step $h \in [H]$, we use $f_h = (o_h, a_h, \cdots, o_{H-1}, a_{H-1}, o_H) \in \mathcal{F}_h := \prod_{h'=h}^{H-1}(\mathcal{O}_{h'} \times \mathcal{A}) \times \mathcal{O}_H$ to denote the future after step h. For future f_h, we define $\mathbf{u}(f_h) \in \mathbb{R}^{|\mathcal{S}_h|}$ as its outcome vector where $[\mathbf{u}(f_h)]_i = \mathbb{P}^{\pi_b}(f_h \mid s_h = i)$. Then we define the outcome matrix $U_{\mathcal{F},h} \in \mathbb{R}^{|\mathcal{F}_h|\times|\mathcal{S}_h|}$ where the row indexed by f_h is $\mathbf{u}(f_h)$. Note that the outcome matrix $U_{\mathcal{F},h}$ depends on the behavior policy π_b, which we omit in the notation.

Off-policy Evaluation (OPE). In OPE, we aim to use an offline dataset collected by a behavior policy π_b to estimate the expected cumulative return of a target policy π_e, which is $J(\pi_e)$. The dataset $\mathcal{D}$ consists of n data trajectories $\{(o_1^{(i)}, a_1^{(i)}, r_1^{(i)}, \ldots, o_H^{(i)}, a_H^{(i)}, r_H^{(i)}) : i \in [n]\}$. Without loss of generality, we assume the reward function R is known and we need to learn the emission and transition dynamics. In this paper, we focus on evaluating history-dependent π_e, which is a more general setting compared to the memoryless π_e considered in previous works (Uehara et al., 2023a; Zhang & Jiang, 2024). Therefore, the action a_h depends on both the history τ_{h-1} and the observation o_h. Throughout the paper, we make the following assumption:

Assumption A. We assume $\pi_b(a_h \mid \tau_{h-1}, o_h)$ is known and $\max_{h,\tau_{h-1},o_h,a_h} \frac{1}{\pi_b(a_h|\tau_{h-1},o_h)} \leq C_{\mathcal{A}}$.

The parameter $C_{\mathcal{A}}$ will not always show up in our upper bounds due to its looseness,[3] but it is a useful relaxation for making meaningful comparisons across different settings of interest in Table 1.

Learning Goal. We consider rich and large observation spaces and want to avoid paying explicit dependence on O in the sample complexity. One way of achieving so is to use importance sampling (Precup, 2000), which avoids O but pays an exponential-in-horizon quantity $O({C_{\mathcal{A}}}^H)$ instead. We follow the setting of Uehara et al. (2023a); Zhang & Jiang (2024) and use function approximation to

[2] The belief state is determined by the environment dynamics and is independent of the policy. Our definition is the same as in Uehara et al. (2023a); Zhang & Jiang (2024); this is slightly different from the usual definition that also includes o_{h+1} as a conditional variable.

[3] See discussion of uniform vs. policy-specific coverage in Section 3.1; the policy-specific counterpart of Assumption A is to bound $\pi_e(a_h \mid \tau_{h-1}, o_h)/\pi_b(a_h \mid \tau_{h-1}, o_h)$.

avoid both O and exponential-in-H, by assuming that the learner has access to a realizable model class $\mathcal{M}$.[4] More formally, we make the following assumption throughout (except in Section 5):

Assumption B (Realizability)**.** We assume that the learner is given $\mathcal{M}$, a class of POMDPs with the same $\mathcal{S}, \mathcal{A}, \mathcal{O}, H, R$ components as $M^\star$. Furthermore, $M^\star \in \mathcal{M}$.

Our goal is to achieve sample complexity bounds polynomial in $\log|\mathcal{M}|, H$ as well as appropriate coverage parameters introduced in the next section.

3 Information-theoretic Hardness of Model-free Algorithms

We start by introducing the key assumptions that enable efficient OPE of *memoryless* policies in prior works. Then, in Section 3.2, we demonstrate via a lower bound that evaluating history-dependent π_e is information-theoretically hard for *any* model-free algorithm (Theorem 3), motivating the investigation of model-based algorithms in Section 4.

3.1 Key Assumptions and Existing Results for Memoryless π_e

The prior work of Zhang & Jiang (2024) identify two key assumptions for OPE of memoryless policies in POMDPs, as introduced below.

Assumption C (Uniform Belief Coverage)**.** Define

$$\Sigma_{\mathcal{H},h} = \mathbb{E}_{\pi_b}[\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\mathbf{b}_{\mathcal{S}}(\tau_{h-1})^\top].$$

Assume that $1/\sigma_{\min}(\Sigma_{\mathcal{H},h}) \le C_{\mathcal{H}}, \forall h \in [H-1]$ for some $C_{\mathcal{H}} < \infty$.

Assumption D (Multi-step Outcome Revealing)**.** Define[5]

$$\Sigma_{\mathcal{F},h} := U_{\mathcal{F},h}^\top Z_h^{-1} U_{\mathcal{F},h}, \quad \text{where} \quad Z_h := \mathrm{diag}(U_{\mathcal{F},h}\mathbf{1}_{\mathcal{S}_h}).$$

Here $\mathrm{diag}(\cdot)$ is a diagonal matrix with its diagonal being the input vector, and $\mathbf{1}_{\mathcal{S}_h}$ is the all-1 vector with dimension $|\mathcal{S}_h|$. We assume that $\|\Sigma_{\mathcal{F},h}^{-1}\|_1 \le C_{\mathcal{F}},\ \forall h \in [H-1]$ for some $C_{\mathcal{F}} < \infty$.

Intuitively, Assumption C states that the belief vector $\mathbf{b}_{\mathcal{S}}(\tau_{h-1})$ spans all directions of $\mathbb{R}^{|\mathcal{S}_h|}$ when τ_{h-1} is generated with π_b, and similar "linear coverage" conditions are also found in the linear MDP literature (Jin et al., 2021; Xiong et al., 2022a).Assumption D is an analogous condition but stated for the future after step h, and can be interpreted as that the future f_h can nontrivially predict the latent state s_h (thus the term "revealing"). In fact, when f_h can deterministically predict s_h, Assumption D holds with $\Sigma_{\mathcal{F},h} = \mathbf{I}$ (the identity matrix) and $C_{\mathcal{F}} = 1$.

Uniform vs. Policy-specific Coverage Both assumptions above are only properties of the behavior policy π_b, whereas the original conditions of Zhang & Jiang (2024) are tighter versions that account for the properties of π_e as well. For example, their belief coverage does not require covering all directions of $\mathbb{R}^{|\mathcal{S}_h|}$, but only the direction of $[d_h^{\pi_e}(s_h = s)]_{s\in\mathcal{S}_h}$. We call the former *uniform coverage* (since it works for all π_e) and the latter *policy-specific coverage* (which is generally tighter). Our positive results in Section 4 also depend on policy-specific coverage parameters, but they are different from the ones in Zhang & Jiang (2024) due to the technical differences between model-free and model-based analyses. Moreover, some definitions (e.g., outcome coverage in Zhang & Jiang (2024), which is the policy-specific version of our Assumption D) do not extend to history-dependent π_e. Therefore, we introduce the uniform version of the assumptions here for a clean and fair comparison.

We now state the result of Zhang & Jiang (2024) in our setup (see Footnote 4 for how we invoke their model-free algorithms and analyses in our model-based setup), with proof in Appendix A.

[4] The algorithms in Uehara et al. (2023a); Zhang & Jiang (2024) are model-free and require value function and Bellman error classes; in Appendix A we show how these classes can be constructed automatically from $\mathcal{M}$, which is a common practice when comparing across model-based and model-free algorithms (Chen & Jiang, 2019; Sun et al., 2019).

[5] The definitions here are based on $U_{\mathcal{F},h}$, calculated using the dynamcis of $M^\star$. Later we will also be interested in variant of this assumption, where all quantities are replaced by their counterparts calculated using the dynamics of some candidate model $M \in \mathcal{M}$.

Theorem 1 (Corollary of Theorem 7 of Zhang & Jiang (2024))**.** *Under Assumptions C and D, assume that π_b and π_e are memoryless, there exists an algorithm (see Appendix A) such that, with probability at least $1-\delta$,[6] $|J(\pi_e) - \widehat{J}(\pi_e)| \leq \epsilon$ with a sample complexity $n = \mathrm{poly}(H, \log(|\mathcal{M}|/\delta), \epsilon, C_\mathcal{A}, C_\mathcal{F}, C_\mathcal{H})$, where $\widehat{J}(\pi_e)$ is the algorithm's estimation of $J(\pi_e)$.*

3.2 Hardness Results for History-dependent π_e

We now show that Theorem 1 cannot hold if we make π_b and π_e general history-dependent policies. In fact, the hardness is not specific to the algorithms of Uehara et al. (2023a); Zhang & Jiang (2024), but applies generally to a very broad range of *model-free* algorithms, as defined below.

Definition 2 (Model-free algorithm)**.** We say that an algorithm for OPE in POMDPs is *model-free*, if it only queries the action distribution of π_e on trajectories observed in the offline dataset, i.e., the only information known about π_e is $\pi_e(\cdot|o_1^{(i)}, a_1^{(i)}, \ldots o_h^{(i)})$ for all $h \in [H], i \in [n]$.

This definition is satisfied by the algorithms of Uehara et al. (2023a); Zhang & Jiang (2024). When it is specialized to the MDP setting, it is also satisfied by most standard algorithms considered model-free, such as importance sampling (Precup, 2000) and Fitted-Q Evaluation (Ernst et al., 2005; Le et al., 2019; Voloshin et al., 2019). The spirit of focusing on how algorithms access and process information in the input is consistent with prior works that show model-free and model-based separation from a learning-theoretic perspective (Chen & Jiang, 2019; Sun et al., 2019).

Moreover, the hardness still holds even if we make the problem easier in two aspects:

1. We keep π_b memoryless (the more general setting is that π_b is also history-dependent just as π_e).
2. Assumption D states that (multi-step) future can reveal the latent state s_h. The online POMDP literature has a related and stronger assumption that the immediate observation o_h plays the same role (Liu et al., 2022a):

Assumption E (Single-step Outcome Revealing)**.** Define

$$\Sigma_{\mathcal{O},h} := \mathbb{O}_h^\top W_h^{-1} \mathbb{O}_h, \quad \text{where} \quad W_h := \mathrm{diag}(\mathbb{O}_h \mathbf{1}_{\mathcal{S}_h}).$$

We assume $\|\Sigma_{\mathcal{O},h}^{-1}\|_1 \leq C_\mathcal{O}\ \forall h \in [H-1]$ for some $C_\mathcal{O} < \infty$.

The interpretation of this assumption is similar to multi-step revealing (Assumption D, but requires that the immediate observation o_h (instead of the entire future as in multi-step revealing) is sufficient for making nontrivial predictions of s_h. Variants of this assumption have been proposed in online learning in POMDPs (Liu et al., 2022a), but the original version has a poor scalability w.r.t. the number of observations, as pointed out by Chen et al. (2022); Zhang & Jiang (2024); see Example 1 of Zhang & Jiang (2024) for further discussions. Our Assumption 4 fixes the issue by using a similar inverse-weighting scheme as Assumption 5 and enjoys better scaling with the size of the observation space. For example, when o_h can uniquely determine s_h, $\Sigma_{\mathcal{O},h} = \mathbf{I}$ and is independent of the number of observations.

Theorem 3 (Information-theoretic hardness of model-free algorithms)**.** *In the same setup as Theorem 1, if we allow π_e to be history-dependent (but π_b is still memoryless), no model-free algorithm can achieve the polynomial sample complexity guarantee. This still holds even if we replace $C_\mathcal{F}$ with the single-step revealing parameter $C_\mathcal{O}$.*

Proof. Consider the following POMDP: for $1 \leq h \leq H-1$, there is only one state $s_{h,1}$ (thus transition before $h = H-1$ is a trivial chain), and $C_\mathcal{H} = 1$. At the last step H, there are two states $s_{H,L}$ and $s_{H,R}$. The emission is identity, i.e., $o_h = s_h$, meaning that the POMDP is also a MDP and $C_\mathcal{O} = C_\mathcal{F} = 1$. There are two actions, L and R. At step $H-1$, taking L and R transitions to $s_{H,L}$ and $s_{H,R}$, respectively. The agent only receives a reward of 1 upon reaching state $s_{H,L}$. The model is known so $|\mathcal{M}| = 1$. The behavior policy π_b is uniformly random, implying $C_\mathcal{A} = 2$. Consider two target policies, π_1 and π_2. Both policies take action L if historical actions do not include R for

[6] Here we assume that the algorithm has knowledge of the value of $C_\mathcal{F}$ and $C_\mathcal{H}$. The analyses can be easily adapted to the case where the precise value of $C_\mathcal{F}$ (or $C_\mathcal{H}$) is unknown but a nontrivial upper bound (e.g., a constant multiple of the value) is known.

$1 \le h \le H-2$. At step $H-1$, π_1 takes action L if all previous actions are L, while π_2 takes action R if all previous actions are L. Under any history that includes at least one R action, both policies yield the same action, and the concrete choice does not matter.[7]

$J(\pi_1) = 1$ and $J(\pi_2) = 0$, as running π_1 produces an all-L sequence, and running π_2 produces $L, \ldots, L, R$. However, since the two policies have identical action choices under all other action sequences, a model-free algorithm can only tell them apart if the sequence $L, \ldots, L$ of length $H-1$ is observed in the offline data, which only happens with a negligible $O(1/2^H)$ probability. With overwhelming probability, π_1 and π_2 will look identical to the algorithm but their $J(\cdot)$ differs by a constant of 1, so no algorithm can predict them well simultaneously up to $\epsilon = 1/2$ accuracy, unless it is given $\Omega(2^H)$ samples to observe the $L, \ldots, L$ sequence with nontrivial probability. However, in this problem, $C_{\mathcal{O}}$, $C_{\mathcal{F}}$, $C_{\mathcal{H}}$, $C_{\mathcal{A}}$, ϵ, $|\mathcal{M}|$ are all constants, so the sample complexity in Theorem 1 is $\text{poly}(H)$, which cannot explain away the exponential in $\Omega(2^H)$. Thus we conclude that no model-free algorithm can achieve Theorem 1's guarantee when π_e is allowed to be history dependent. □

The counter-intuitive part of the construction is that the model is fully known ($\mathcal{M} = \{M^\star\}$), and the hardness fully comes from the fact that model-free algorithms have limited access to π_e and cannot distinguish between two history-dependent policies with very different returns from data. In fact, if we remove the model-free restriction, the problem instance is trivial to solve as we can calculate $J_{M^\star}(\pi_e)$ by rolling out π_e in $M^\star$.

4 Positive Results for Model-based Algorithms

The negative result from the previous section naturally motivates the investigation of model-based algorithms, i.e., those that do not respect the restriction of Definition 2. Model-based algorithms fit the POMDP dynamics from data and use the learned model to evaluate π_e; the latter step can often be achieved by rolling out trajectories in the learned model, where we need to query π_e on *synthetic* trajectories generated by the learned model, thus violating Definition 2 and potentially circumventing the hardness in Theorem 3.

Concretely, we consider a very natural algorithm based on Maximum Likelihood Estimation (MLE). It is, in fact, quite surprising that such a simple algorithm has not been analyzed under relatively general assumptions for OPE in POMDPs. The algorithm is as follows:

- **Pre-filtering:** We construct $\mathcal{M}' \subset \mathcal{M}$ where all models that violate Assumption E (for Section 4.1 where we assume single-step revealing in the guarantee) or D (for Section 4.2 where we assume multi-step revealing) are excluded from $\mathcal{M}'$.[8] Such a filtering step is common in the POMDP learning literature (Liu et al., 2022a), and in Section 5 we show why MLE estimation requires this as a pre-processing step and will fail otherwise.
- **MLE:** Let $\mathbb{P}_M^\pi$ stands for the probabilities under policy π and model M. The model is learned as

$$\widehat{M} = \max_{M \in \mathcal{M}'} \textstyle\sum_{i=1}^n \log \mathbb{P}_M^\pi(\tau_H^{(i)}). \tag{1}$$

- **Prediction:** We use the expected return of π_e in $\widehat{M}$, $J_{\widehat{M}}(\pi_e)$, as our estimation for $J(\pi_e)$.

4.1 Guarantee under Single-step Outcome Revealing

We first show that the model-based algorithm enjoys a polynomial guarantee when we assume the stronger single-step outcome revealing (Assumption E) instead of its multi-step version (Assumption D). Recall that even with this stronger revealing assumption the model-free algorithms still cannot provide polynomial guarantees.

[7] Rolling out π_1 and π_2 leads to deterministic outcomes, which are characteristics of open-loop policies (i.e., action only depends on h and is independent of (τ_{h-1}, o_h)). However, π_1 and π_2 cannot be open-loop simultaneously, otherwise they cannot have identical action choices on histories that include R.

[8] This requires the algorithm to have knowledge of the value or some nontrivial upper bound of $C_{\mathcal{O}}$ (or $C_{\mathcal{F}}$). Such knowledge is also required for Theorem 1.

Theorem 4. *Given a realizable model class $\mathcal{M}$, let model $\widehat{M}$ be the MLE within $\mathcal{M}'$ using the dataset $\mathcal{D}$. Under Assumption E, with probability at least $1-\delta$, we have*

$$\left|J(\pi_e) - J_{\widehat{M}}(\pi_e)\right| \leq \mathcal{O}\left(H^2 C_{\mathcal{O}}^2 C_{\textit{eff},1}\sqrt{\frac{\log\frac{|\mathcal{M}|}{\delta}}{n}}\right).$$

where

$$C_{\textit{eff},1} := \max_{h\in[H-1]} \frac{\sum_{o_h,a_h}\mathbb{E}_{\pi_e}\left[\pi_e(a_h \mid \tau_{h-1}, o_h)\|(\widehat{\mathbf{B}}_h - \mathbf{B}_h)\mathbb{O}_h \mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_1\right]}{\sum_{o_h,a_h}\mathbb{E}_{\pi_b}\left[\pi_b(a_h \mid \tau_{h-1}, o_h)\|(\widehat{\mathbf{B}}_h - \mathbf{B}_h)\mathbb{O}_h \mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_1\right]},$$

Here $\mathbf{B}_h = \mathbf{B}_h(o_h, a_h)$ is the observable-operator parameterization of $M^\star$ under single-step revealing (see Appendix C), and $\widehat{\mathbf{B}}_h$ is defined similarly for $\widehat{M}$. Furthermore, when Assumptions A and C hold, we have

$$C_{\textit{eff},1} \leq C_{\mathcal{A}} C_{\mathcal{H}}.$$

The proof is deferred to Appendix D.1. To the best of our knowledge, this is the first polynomial sample-complexity bound for model-based OPE of history-dependent target policies in POMDPs with large observation spaces. In addition to the quantities that have already been introduced, the bound depends on a new coverage parameter, $C_{\text{eff},1}$, which we show can be upper bounded by $C_{\mathcal{A}}C_{\mathcal{H}}$. With this relaxation, Theorem 4 implies a formal poly$((H, \log(|\mathcal{M}|/\delta), \epsilon, C_{\mathcal{A}}, C_{\mathcal{F}}, C_{\mathcal{H}})$ sample complexity, in a setting where model-free OPE provably cannot handle (Theorem 3).

However, as mentioned in Section 3.2, the $C_{\mathcal{A}}C_{\mathcal{H}}$ bound is relaxing $C_{\text{eff},1}$ to a "uniform coverage" parameter, whereas $C_{\text{eff},1}$ itself is policy-sepcific (note the dependence on π_e on the numerator) and can be much tighter. For starters, when $\pi_e = \pi_b$, it is obvious that $C_{\text{eff},1} = 1$. More generally, $C_{\text{eff},1}$ captures the discrepancy between the model estimation error under the distributions induced by π_e and π_b. Unlike the commonly used L_2 norm of Bellman error in offline MDPs (Xie et al., 2021), we use the L_1 norm of the estimation error, which better leverages the property of the belief state vector ($\|\mathbf{b}_{\mathcal{S}}(\tau_h)\|_1 = 1$). The advantage of L_1 over L_2 is also discussed in Zhang & Jiang (2024), where they note that L_1/L_∞ Hölder's inequality can help reduce the dependence on S in the analysis. In the special case of the MDP setting (i.e., $s_h = o_h$ and $\mathbb{O}_h \equiv \mathbf{I}$), we have (see Appendix B)

$$C_{\text{eff},1} \leq \max_h \max_{s_h,a_h} \frac{d_h^{\pi_e}(s_h,a_h)}{d_h^{\pi_b}(s_h,a_h)}.$$

This recovers a coverage coefficient in the form of state-action density ratio, which is commonly used in the offline MDP literature (Munos, 2007; Antos et al., 2008; Chen & Jiang, 2019; Jiang & Xie, 2024).

4.2 Guarantee under Multi-step Outcome Revealing

Theorem 4 relies on Assumption E, which requires that the observation reveals sufficient information about the latent state. However, this assumption may not hold in more complex partially observable environments. We now show that the model-based algorithm can also enjoy polynomial guarantees under the weaker multi-step revealing condition (Assumption D), if we impose that the behavior policy π_b is memoryless (the target policy π_e is still history dependent).

Theorem 5. *Given a realizable model class $\mathcal{M}$, let model $\widehat{M}$ be the MLE within $\mathcal{M}'$ using dataset $\mathcal{D}$. Suppose π_b is memoryless and Assumption D hold, with probability at least $1-\delta$, we have*

$$\left|J(\pi_e) - J_{\widehat{M}}(\pi_e)\right| \leq \mathcal{O}\left(H^2 C_{\mathcal{F}}^2 C_{\textit{eff},m}\sqrt{\frac{\log\frac{|\mathcal{M}|}{\delta}}{n}}\right).$$

where

$$C_{\textit{eff},m} := \max_{h\in[H-1]} \frac{\sum_{o_h,a_h}\mathbb{E}_{\pi_e}\left[\pi_e(a_h \mid \tau_{h-1}, o_h)\|(\widehat{\mathbf{B}}_h - \mathbf{B}_h)U_{\mathcal{F},h} \mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_1\right]}{\sum_{o_h,a_h}\mathbb{E}_{\pi_b}\left[\pi_b(a_h \mid \tau_{h-1}, o_h)\|(\widehat{\mathbf{B}}_h - \mathbf{B}_h)U_{\mathcal{F},h} \mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_1\right]}, \tag{2}$$

where $\mathbf{B}_h = \mathbf{B}_h(o_h, a_h)$ is the observable-operator parameterization of $M^\star$ under multi-step revealing (Appendix C). Furthermore, similar to the single-step case, when Assumption C holds we can upper bound $C_{\text{eff},m}$ as

$$C_{\text{eff},m} \leq C_{\mathcal{A}} C_{\mathcal{H}}.$$

The proof is deferred to Appendix D.2. Our theorem shows that when the behavior policy π_b is memoryless, only polynomial samples are required to evaluate the target policy π_e under Assumption D. The structure of the bound is similar to that of Theorem 4, except that it replaces $C_{\mathcal{O}}$ with its multi-step counterpart $C_{\mathcal{F}}$.

Interestingly, our results reveal that a memoryless behavior policy can accurately evaluate history-dependent target policies when certain coverage conditions are satisfied. This is somewhat counter-intuitive, as one might expect that only history-dependent behavior policies could evaluate history-dependent target policies. However, this conclusion aligns with findings in latent MDPs (Kwon et al., 2024), which are a special case of POMDPs. This insight encourages the use of memoryless policies with good coverage for online exploration, consistent with many online algorithms that employ uniform exploration policies (Efroni et al., 2022; Liu et al., 2022a; Uehara et al., 2022a; Guo et al., 2023b).

5 State-Space Mismatch and Model Misspecification

So far we assume the agent is given a realizable model class with the ground-truth latent-state space size $|\mathcal{S}|$. Although the knowledge of $|\mathcal{S}|$ is commonly assumed in both OPE in POMDPs (Uehara et al., 2023a; Zhang & Jiang, 2024) and online learning for POMDPs (Jin et al., 2020; Liu et al., 2022a), obtaining this information in practice can be challenging, as the latent state space is an ungrounded object. A natural concern is that of state-space misspecification (Kulesza et al., 2014; 2015): what if the state space of models in $\mathcal{M}$ (denoted as $\widehat{\mathcal{S}}$ in this section) is different from that of $M^\star$? This question has largely eluded the recent literature on RL in POMDPs, and below we show that taking it seriously reveals interesting insights.

One immediate consequence of $|\widehat{\mathcal{S}}| \neq |\mathcal{S}|$ is that the model realizability assumption $M^\star \in \mathcal{M}$ no longer makes any sense. A natural rescue is to note that it is still possible to have models in $\mathcal{M}$ that induce the same *observable process* as $M^\star$, that is, they always generate the same distribution of observation-action sequences given any policy. For example, if $|\widehat{\mathcal{S}}| > |\mathcal{S}|$, we can still hope this weaker notion of realizability (based on observable equivalence) holds, since one can simply add dummy latent states to $M^\star$ to create an equivalent model with larger latent-state spaces. More concretely,

Definition 6. We say that $\mathcal{M}$ satisfies *observable-equivalent* realizability, if there exists $M \in \mathcal{M}$, such that $P_M(o_{h+1}|o_1, a_1, \ldots, o_h, a_h)$ is the same as $P_{M^\star}(o_{h+1}|o_1, a_1, \ldots, o_h, a_h)$ for all h and $o_1, a_1, \ldots, o_{h+1}$.

We now show a somewhat counter-intuitive result: replacing $M^\star \in \mathcal{M}$ with observable-equivalent realizability will break the previous guarantee (e.g., Theorem 4).

Theorem 7. *Take the same setting as Theorem 4, with model realizability (Assumption B) replaced by the weaker observable-equivalent realizability (Definition 6). The model-based algorithm, either with or without the pre-filtering step (Section 4), cannot achieve the polynomial guarantee in Theorem 4.*

Proof. The action space is $\{L, R\}$ and rewards are only received at final step H. $\mathcal{M} = \{M_1, M_2\}$, i.e., there are only two candidate models in $\mathcal{M}$. All the models including $M^\star$ share the same observation space, which contains only one observation for $h \in [H-1]$ and two observations at step H. M_1 and M_2 have the same state space as $M^\star$ at the final step H: s_{good} and s_{bad}, state s_{good} always generates o_{good} with reward 1 and state s_{bad} always generates o_{bad} with reward 0. In M_1, the state space at step $h \in [H-1]$ is $\{L, R\}^{h-1}$. For $h \in [H-2]$, any state-action pair (s_h, a_h) deterministically transitions to the next state $s_{h+1} = (s_h, a_h)$, where a_h is appended to s_h. At step $H-1$, all states transition to s_{good} when action L is taken; otherwise, they transition to s_{bad}. M_2 shares the same state space and dynamics as M_1, except that at step $H-1$, the all-R state will transition to s_{good} if action R is taken. The ground-truth model $M^\star$ has only one state at each step

$h \in [H-1]$. At step $H-1$, it transitions to s_{good} if action L is taken; otherwise, it transitions to s_{bad}. We can verify that M_1 produces the same observable process as $M^\star$. For $M^\star$, since there is only one state and one observation for $h \in [H-1]$, we have $C_{\mathcal{O}} = C_{\mathcal{H}} = 1$. The behavior policy π_b is uniformly random with $C_{\mathcal{A}} = 2$, while the target policy π_e always takes action R.

We first consider the case where the pre-filtering step is skipped, so both M_1 and M_2 are considered for MLE. To estimate π_e accurately, the algorithm needs to output $\widehat{M} = M_1$ (M_2's prediction is wrong by a constant), but the empirical MLE losses (Eq.(1)) of M_1 and M_2 are always identical unless the all-R action sequence is contained in the dataset . Since π_b is uniformly random, $\Omega(2^H)$ samples are needed to include this sequence with nontrivial probability. Therefore, with only poly(H) samples, MLE cannot distinguish between M_1 and M_2, and hence cannot accurately estimate $J(\pi_e)$.

If the algorithm does perform the pre-filtering step, note that $C_{\mathcal{O}} = \infty$ for both models as $|\mathcal{S}_h| > |\mathcal{O}_h|$ for $h > 1$ and $\Sigma_{\mathcal{O},h}$ is non-invertible, so both models will be eliminated and the algorithm cannot be executed. □

In the model-based algorithm in Section 4, we assume revealing assumptions (Assumptions E or D) on $M^\star$ and perform pre-filtering. The negative result here sheds light on this and reveals the underlying reason: what we really need is not $M^\star$, but the learned model $\widehat{M}$ to satisfy revealing properties. In the proof of Theorem 7, no models in $\mathcal{M}$ satisfy single-step revealing (Assumption E) which leads to poor sample efficiency. Inspired by this observation, we present the following result that handles state-space mismatch properly. We assume $|\widehat{\mathcal{S}}|$ can be either smaller or greater than $|\mathcal{S}|$, and an approximate version of observable-equivalent realizability.

Theorem 8. *Given a model class $\mathcal{M}$ such that each $M \in \mathcal{M}$ satisfies Assumption D. Let model $\widehat{M}$ be the MLE within $\mathcal{M}$ using dataset $\mathcal{D}$. Suppose π_b is memoryless, with probability at least $1-\delta$,*

$$\left|J(\pi_e) - J_{\widehat{M}}(\pi_e)\right| \le \mathcal{O}\left(H^2 C_{\mathcal{F}}^2 C_{\textit{eff},m} \sqrt{\frac{\log \frac{|\mathcal{M}|}{\delta}}{n} + \varepsilon_{\textit{approx}}}\right),$$

where $C_{\textit{eff},m}$ is defined in Eq. (2) and $\varepsilon_{\textit{approx}}$ is the approximation error of $\mathcal{M}$:

$$\varepsilon_{\textit{approx}} := \min_{M \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \left(\log \mathbb{P}^{\pi}_{M^\star}(\tau_H^{(i)}) - \log \mathbb{P}^{\pi}_{M}(\tau_H^{(i)})\right).$$

The proof is deferred to Appendix E. Here we assume that all $M \in \mathcal{M}$ satisfies Assumption D. It is actually fine if $M^\star$ itself does not satisfy the assumption, as long as one of models $M \in \mathcal{M}$ is observable-equivalent to $M^\star$ up to approximation error $\varepsilon_{\text{approx}}$.

6 Related Works

OPE in POMDPs. There are two main directions in the study of OPE in POMDPs: OPE in confounded POMDPs and OPE in unconfounded POMDPs. OPE in confounded POMDPs (Zhang & Bareinboim, 2016; Namkoong et al., 2020; Tennenholtz et al., 2020; Nair & Jiang, 2021; Shi et al., 2022; Xu et al., 2023; Bennett & Kallus, 2024) assume that the actions from the behavior policy depend only on the latent state. As a result, these methods are inapplicable to our unconfounded setting, where the latent state is unobservable, and the behavior policy depends solely on the observations and actions. A line of research (Hu & Wager, 2023; Uehara et al., 2023a; Zhang & Jiang, 2024) investigates the same unconfounded setting as ours. Hu & Wager (2023) employ multi-step importance sampling, which leads to an exponential dependence on the horizon length. Uehara et al. (2023a); Zhang & Jiang (2024) use model-free methods that estimate the future-dependent value functions of the target policy, but their approaches focus on memoryless target policies or policies with limited memory. In this paper, we leverage model-based methods and provide polynomial sample complexity bounds for history-dependent target policies.

OPE in MDPs. OPE is an important problem in MDPs and many works have studied it, including importance sampling methods (Precup, 2000; Li et al., 2011), marginalized importance sampling approaches (Liu et al., 2018; Xie et al., 2019; Kallus & Uehara, 2020; Katdare et al., 2023), doubly robust estimators (Dudík et al., 2011; Thomas & Brunskill, 2016; Jiang & Li, 2016; Farajtabar et al., 2018; Xu et al., 2021) and model-based methods (Eysenbach et al., 2020; Voloshin et al., 2021; Yin et al., 2021). However, most of these methods (except for importance sampling) require the Markovian property of the environment. Applying these techniques to POMDPs either does not work or results in an exponential dependence on the horizon length. Xie et al. (2019) demonstrated that state-action coverage is sufficient to achieve polynomial sample complexity in MDPs. However, as shown by Kwon et al. (2024), latent state coverage is insufficient in POMDPs due to their partial observability. In this paper, we consider coverage conditions for both histories and outcomes, a direction that is also explored in Zhang & Jiang (2024) for model-free methods.

Online Learning in POMDPs. Online learning algorithms in POMDPs have been extensively studied. Krishnamurthy et al. (2016) show that the lower bound on the sample complexity of learning general POMDPs is exponential in the horizon, and later works have focused on circumventing the hardness with additional assumptions. Uehara et al. (2023b) propose provably efficient algorithms for POMDPs with deterministic transition dynamics. Guo et al. (2016); Azizzadenesheli et al. (2016); Xiong et al. (2022b) assume the environment satisfies the reachability assumption or that exploratory data is available. There are also many works studying sub-classes of POMDPs, including latent MDPs (Kwon et al., 2021; 2024), decodable POMDPs (Krishnamurthy et al., 2016; Jiang et al., 2017; Du et al., 2019; Efroni et al., 2022), weakly revealing POMDPs (Jin et al., 2020; Liu et al., 2022a;b), low-rank POMDPs (Wang et al., 2022; Guo et al., 2023b), POMDPs with hindsight observability (Lee et al., 2023; Guo et al., 2023a) and observable POMDPs (Golowich et al., 2022a;b), except for the model-free algorithm of Uehara et al. (2022a) based on the future-dependent value function framework (Uehara et al., 2023a).

7 Conclusion

We studied OPE of history-dependent target policies in POMDPs with large observation spaces, and showed provable separation between model-free and model-based methods in several settings. A major open problem is whether model-based algorithm can handle history-dependent π_b and multi-step revealing (the "?" mark in Table 1), and resolving this question will provide a more comprehensive picture of the landscape of OPE in POMDPs.

References

András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71:89–129, 2008.

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

Andrew Bennett and Nathan Kallus. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *Operations Research*, 72(3):1071–1086, 2024.

Léon Bottou, Jonas Peters, Joaquin Quiñonero-Candela, Denis X Charles, D Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Simard, and Ed Snelson. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14 (11), 2013.

Olivier Chapelle, Eren Manavoglu, and Romer Rosales. Simple and scalable response prediction for display advertising. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):1–34, 2014.

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

Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Simon Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudik, and John Langford. Provably efficient rl with rich observations via latent state decoding. In *International Conference on Machine Learning*, pp. 1665–1674. PMLR, 2019.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Yonathan Efroni, Chi Jin, Akshay Krishnamurthy, and Sobhan Miryoosefi. Provable reinforcement learning with a short-term memory. In *International Conference on Machine Learning*, pp. 5832–5850. PMLR, 2022.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pp. 667–672. IEEE, 2006.

Benjamin Eysenbach, Swapnil Asawa, Shreyas Chaudhari, Sergey Levine, and Ruslan Salakhutdinov. Off-dynamics reinforcement learning: Training for transfer with domain classifiers. *arXiv preprint arXiv:2006.13916*, 2020.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. In *International Conference on Machine Learning*, pp. 1447–1456. PMLR, 2018.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Learning in observable pomdps, without computationally intractable oracles. *Advances in neural information processing systems*, 35: 1458–1473, 2022a.

Noah Golowich, Ankur Moitra, and Dhruv Rohatgi. Planning in observable pomdps in quasipolynomial time. *arXiv preprint arXiv:2201.04735*, 2022b.

Jiacheng Guo, Minshuo Chen, Huan Wang, Caiming Xiong, Mengdi Wang, and Yu Bai. Sample-efficient learning of pomdps with multiple observations in hindsight. *arXiv preprint arXiv:2307.02884*, 2023a.

Jiacheng Guo, Zihao Li, Huazheng Wang, Mengdi Wang, Zhuoran Yang, and Xuezhou Zhang. Provably efficient representation learning with tractable planning in low-rank pomdp. In *International Conference on Machine Learning*, pp. 11967–11997. PMLR, 2023b.

Zhaohan Daniel Guo, Shayan Doroudi, and Emma Brunskill. A pac rl algorithm for episodic pomdps. In *Artificial Intelligence and Statistics*, pp. 510–518. PMLR, 2016.

Yuchen Hu and Stefan Wager. Off-policy evaluation in partially observed markov decision processes under sequential ignorability. *The Annals of Statistics*, 51(4):1561–1585, 2023.

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

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pp. 652–661. PMLR, 2016.

Nan Jiang and Tengyang Xie. Offline reinforcement learning in large state spaces: Algorithms and guarantees. 2024.

Nan Jiang, Akshay Krishnamurthy, Alekh Agarwal, John Langford, and Robert E Schapire. Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pp. 1704–1713. PMLR, 2017.

Chi Jin, Sham Kakade, Akshay Krishnamurthy, and Qinghua Liu. Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 33: 18530–18539, 2020.

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

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

Pulkit Katdare, Nan Jiang, and Katherine Rose Driggs-Campbell. Marginalized importance sampling for off-environment policy evaluation. In *Conference on Robot Learning*, pp. 3778–3788. PMLR, 2023.

Akshay Krishnamurthy, Alekh Agarwal, and John Langford. PAC reinforcement learning with rich observations. In *Advances in Neural Information Processing Systems*, 2016.

Alex Kulesza, N Raj Rao, and Satinder Singh. Low-rank spectral learning. In *Artificial Intelligence and Statistics*, pp. 522–530. PMLR, 2014.

Alex Kulesza, Nan Jiang, and Satinder Singh. Low-rank spectral learning with weighted loss functions. In *Artificial Intelligence and Statistics*, pp. 517–525. PMLR, 2015.

Jeongyeol Kwon, Yonathan Efroni, Constantine Caramanis, and Shie Mannor. Rl for latent mdps: Regret guarantees and a lower bound. *Advances in Neural Information Processing Systems*, 34: 24523–24534, 2021.

Jeongyeol Kwon, Shie Mannor, Constantine Caramanis, and Yonathan Efroni. Rl in latent mdps is tractable: Online guarantees via off-policy evaluation. *arXiv preprint arXiv:2406.01389*, 2024.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pp. 3703–3712. PMLR, 2019.

Jonathan Lee, Alekh Agarwal, Christoph Dann, and Tong Zhang. Learning in pomdps is sample-efficient with hindsight observability. In *International Conference on Machine Learning*, pp. 18733–18773. PMLR, 2023.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 297–306, 2011.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.

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

Qinghua Liu, Csaba Szepesvári, and Chi Jin. Sample-efficient reinforcement learning of partially observable markov games. *Advances in Neural Information Processing Systems*, 35:18296–18308, 2022b.

Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, volume 1077, 2014.

Rémi Munos. Performance bounds in l_p-norm for approximate value iteration. *SIAM journal on control and optimization*, 46(2):541–561, 2007.

Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):331–355, 2003.

Yash Nair and Nan Jiang. A spectral approach to off-policy evaluation for pomdps. *arXiv preprint arXiv:2109.10502*, 2021.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33:18819–18831, 2020.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, pp. 80, 2000.

Chengchun Shi, Masatoshi Uehara, Jiawei Huang, and Nan Jiang. A minimax learning approach to off-policy evaluation in confounded partially observable markov decision processes. In *International Conference on Machine Learning*, pp. 20057–20094. PMLR, 2022.

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

Guy Tennenholtz, Uri Shalit, and Shie Mannor. Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10276–10283, 2020.

Georgios Theocharous, Philip S Thomas, and Mohammad Ghavamzadeh. Ad recommendation systems for life-time value optimization. In *Proceedings of the 24th international conference on world wide web*, pp. 1305–1310, 2015.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 2139–2148. PMLR, 2016.

Masatoshi Uehara, Ayush Sekhari, Jason D Lee, Nathan Kallus, and Wen Sun. Provably efficient reinforcement learning in partially observable dynamical systems. *Advances in Neural Information Processing Systems*, 35:578–592, 2022a.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022b.

Masatoshi Uehara, Haruka Kiyohara, Andrew Bennett, Victor Chernozhukov, Nan Jiang, Nathan Kallus, Chengchun Shi, and Wen Sun. Future-dependent value-based off-policy evaluation in pomdps. *Advances in Neural Information Processing Systems*, 36, 2023a.

Masatoshi Uehara, Ayush Sekhari, Jason D Lee, Nathan Kallus, and Wen Sun. Computationally efficient pac rl in pomdps with latent determinism and conditional embeddings. In *International Conference on Machine Learning*, pp. 34615–34641. PMLR, 2023b.

Cameron Voloshin, Hoang M Le, Nan Jiang, and Yisong Yue. Empirical study of off-policy policy evaluation for reinforcement learning. *arXiv preprint arXiv:1911.06854*, 2019.

Cameron Voloshin, Nan Jiang, and Yisong Yue. Minimax model learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1612–1620. PMLR, 2021.

Lingxiao Wang, Qi Cai, Zhuoran Yang, and Zhaoran Wang. Embed to control partially observed systems: Representation learning with provable sample efficiency. *arXiv preprint arXiv:2205.13476*, 2022.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in neural information processing systems*, 32, 2019.

Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512*, 2022a.

Yi Xiong, Ningyuan Chen, Xuefeng Gao, and Xiang Zhou. Sublinear regret for learning pomdps. *Production and Operations Management*, 31(9):3491–3504, 2022b.

Tengyu Xu, Zhuoran Yang, Zhaoran Wang, and Yingbin Liang. Doubly robust off-policy actor-critic: Convergence and optimality. In *International Conference on Machine Learning*, pp. 11581–11591. PMLR, 2021.

Yang Xu, Jin Zhu, Chengchun Shi, Shikai Luo, and Rui Song. An instrumental variable approach to confounded off-policy evaluation. In *International Conference on Machine Learning*, pp. 38848–38880. PMLR, 2023.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1567–1575. PMLR, 2021.

Junzhe Zhang and Elias Bareinboim. Markov decision processes with unobserved confounders: A causal approach. *Purdue AI Lab, West Lafayette, IN, USA, Tech. Rep*, 2016.

Yuheng Zhang and Nan Jiang. On the curses of future and history in future-dependent value functions for off-policy evaluation. *arXiv preprint arXiv:2402.14703*, 2024.

A Proof of Theorem 1

Proof. We use the algorithm described in Section 3 of Zhang & Jiang (2024). Given the model class $\mathcal{M}$, we construct the function classes $\mathcal{V}$ and Ξ needed by their algorithm. First, we apply the same pre-filtering as in the model-based algorithm to obtain $\mathcal{M}'$. For each $M \in \mathcal{M}'$, we include the corresponding future-dependent value function (FDVF) $V_{\mathcal{F}}$ (their Definition 3) in $\mathcal{V}$. For each model pair $M_1, M_2 \in \mathcal{M}'$, we include $\mathcal{B}^{\mathcal{H}}_{M_1} V_{\mathcal{F},M_2}$ (their Definition 4) in Ξ, where $V_{\mathcal{F},M_2}$ is the FDVF in model M_2 and $\mathcal{B}^{\mathcal{H}}_{M_1}$ is the Bellman residual operator under model M_1. Hence, we have $|\mathcal{V}| \le |\mathcal{M}|$ and $|\Xi| \le |\mathcal{M}|^2$. Meanwhile, their Assumptions 5 and 6 are naturally satisfied. From our Assumption C, their Assumption 11 is also satisfied with $C_{\mathcal{H},2} \le C_{\mathcal{H}}^2$. According to their Lemma 5, we have $\|V_{\mathcal{F}}\|_\infty \le HC_{\mathcal{F}}$. Then, we invoke their Theorem 7 and obtain that w.p. $1-\delta$,

$$|J(\pi_e) - \widehat{J}(\pi_e)| \le \mathcal{O}\left(H^2 C_{\mathcal{F}} C_{\mathcal{H}} \sqrt{\frac{C_{\mathcal{A}} \log(|\mathcal{M}|/\delta)}{n}}\right). \qquad \square$$

B Calculation of the MDP Case

For MDP, we have $o_h = s_h$ and $\mathbb{O}_h \equiv I$. Hence, given τ_{h-1}, s_h, a_h,

$$\begin{aligned}
&\pi_e(a_h \mid \tau_{h-1}, s_h) \left\| \left(\mathbf{B}_h(s_h,a_h) - \widehat{\mathbf{B}}_h(s_h,a_h)\right) \mathbb{O}_h \mathbf{b}_{\mathcal{S}}(\tau_{h-1}) \right\|_1 \\
&= \pi_e(a_h \mid \tau_{h-1}, s_h) \left\| (\mathbb{T}_{h,a_h} - \widehat{\mathbb{T}}_{h,a_h}) \mathrm{diag}(\mathbb{O}_h(s_h \mid \cdot)) \mathbf{b}_{\mathcal{S}}(\tau_{h-1}) \right\|_1 \\
&= \pi_e(a_h \mid \tau_{h-1}, s_h) \sum_{s'} \mathbb{P}_{M^\star}(s_h \mid \tau_{h-1}) \left| \mathbb{P}_{\widehat{M}}(s' \mid s_h, a_h) - \mathbb{P}_{M^\star}(s' \mid s_h, a_h) \right| \\
&= \mathbb{P}^{\pi_e}_{M^\star}(s_h, a_h \mid \tau_{h-1}) \left\| \mathbb{P}_{\widehat{M}}(\cdot \mid s_h, a_h) - \mathbb{P}_{M^\star}(\cdot \mid s_h, a_h) \right\|_1 .
\end{aligned}$$

Fix step h, the numerator in $C_{\text{eff},1}$ becomes

$$\sum_{s_h,a_h} d_h^{\pi_e}(s_h,a_h) \left\| \mathbb{P}_{\widehat{M}}(\cdot \mid s_h, a_h) - \mathbb{P}_{M^\star}(\cdot \mid s_h, a_h) \right\|_1 .$$

Similarly, the denominator is

$$\sum_{s_h,a_h} d_h^{\pi_b}(s_h,a_h) \left\| \mathbb{P}_{\widehat{M}}(\cdot \mid s_h, a_h) - \mathbb{P}_{M^\star}(\cdot \mid s_h, a_h) \right\|_1 .$$

Then we have

$$C_{\text{eff},1} \le \max_h \max_{s_h,a_h} \frac{d_h^{\pi_e}(s_h,a_h)}{d_h^{\pi_b}(s_h,a_h)}.$$

C Observable Operator Models

We introduce the observable operator models (OOMs) (Jaeger, 2000), which (for our purposes) can be viewed as a reparameterization of POMDPs. OOMs have seen wide use in recent POMDP literature (especially in the online setting (Liu et al., 2022a)), and play a crucial role in our model-based analyses.

Single-step Revealing. We first introduce the OOM used in environments satisfying single-step revealing condition. Given model parameters $M^\star = (\mathbb{T}, \mathbb{O}, d_1)$, we write the initial observation distribution $\mathbf{b}_0$ and observable operators $\mathbf{B}$ as

$$\begin{aligned}
\mathbf{b}_0 &= \mathbb{O}_1 d_1 \in \mathbb{R}^O, \\
\mathbf{B}_h(o,a) &= \mathbb{O}_{h+1} \mathbb{T}_{h,a} \mathrm{diag}(\mathbb{O}_h(o \mid \cdot)) \mathbb{O}_h^{\dagger,w} \in \mathbb{R}^{O \times O},
\end{aligned}$$

where $\mathbb{T}_{h,a} \in \mathbb{R}^{|\mathcal{S}_{h+1}| \times |\mathcal{S}_h|}$ with $[\mathbb{T}_{h,a}]_{ij} = \mathbb{T}_h(s_{h+1} = i \mid s_h = j, a_h = a)$. Different from the pseudo-inverse in Liu et al. (2022a), we use the weighted version (Zhang & Jiang, 2024) defined as follows:

$$\mathbb{O}_h^{\dagger,w} = \Sigma_{\mathcal{O},h}^{-1} \mathbb{O}_h^\top W_h^{-1},$$

where $\Sigma_{\mathcal{O},h}$ and W_h are defined in Assumption E. With the OOM representaion, for any trajectory $\tau_h = (o_1, a_1, \cdots, o_h, a_h)$, the probability of generating τ_h with policy π can be written as

$$\mathbb{P}^{\pi}_{M^\star}(\tau_h) = \pi(\tau_h) \cdot \left(\mathbf{e}_{o_h}^\top \mathbf{B}_{h-1}(o_{h-1}, a_{h-1}) \cdots \mathbf{B}_1(o_1, a_1)\mathbf{b}_0\right),$$

where $\pi(\tau_h) = \prod_{h'=1}^{h} \pi(a_h \mid o_h, \tau_{h-1})$ is the action probability of generating τ_h given policy π. Moreover, let $\mathbf{b}(\tau_h) = \left(\prod_{h'=1}^{h} \mathbf{B}_{h'}(o_{h'}, a_{h'})\right) \mathbf{b}_0$, we have the following observation:

$$\mathbf{b}(\tau_h) = \mathbb{P}_{M^\star}(\tau_h)\mathbb{O}_{h+1}\mathbf{b}_{\mathcal{S}}(\tau_h),$$

where $\mathbb{P}_{M^\star}(\tau_h) = \mathbf{e}_{o_h}^\top \mathbf{B}_{h-1}(o_{h-1}, a_{h-1}) \cdots \mathbf{B}_1(o_1, a_1)\mathbf{b}_0$ is the environment probability of generating τ_h. The relationship between $\mathbf{b}(\tau_h)$ and belief state vector $\mathbf{b}_{\mathcal{S}}(\tau_h)$ is crucial for developing coverage conditions on the belief state.

Multi-step Revealing. Next, we introduce the OOM in the multi-step revealing setting. Recall that f_h denotes the future after step h, and $U_{\mathcal{F},h}$ is the outcome matrix, where the row indexed by f_h represents its corresponding outcome vector. The observable operators are defined as

$$\begin{aligned}
&\mathbf{b}_0 = U_{\mathcal{F},1} d_1 \in \mathbb{R}^{|\mathcal{F}_1|}, \\
&\mathbf{B}_h(o, a) = U_{\mathcal{F},h+1}\mathbb{T}_{h,a}\mathrm{diag}(\mathbb{O}_h(o \mid \cdot))U^{\dagger,w}_{\mathcal{F},h} \in \mathbb{R}^{|\mathcal{F}_{h+1}| \times |\mathcal{F}_h|}.
\end{aligned}$$

Here $U^{\dagger,w}_{\mathcal{F},h} = \Sigma^{-1}_{\mathcal{F},h} U^\top_{\mathcal{F},h} Z_h^{-1}$ is the weighted pseudo-inverse and $\Sigma_{\mathcal{F},h}$, Z_h are defined in Assumption D. For any history τ_{h-1} and future f_h, the probability of generating (τ_{h-1}, f_h) with policy π is written as

$$\mathbb{P}^{\pi}_{M^\star}(\tau_{h-1}, f_h) = \pi(\tau_{h-1}) \cdot \left(\mathbf{e}_{f_h}^\top \mathbf{B}_{h-1}(o_{h-1}, a_{h-1}) \cdots \mathbf{B}_1(o_1, a_1)\mathbf{b}_0\right),$$

Similar to the single-step case, we also have the following relation between $\mathbf{b}(\tau_h)$ and $\mathbf{b}_{\mathcal{S}}(\tau_h)$:

$$\mathbf{b}(\tau_h) = \mathbb{P}_{M^\star}(\tau_h)U_{\mathcal{F},h+1}\mathbf{b}_{\mathcal{S}}(\tau_h),$$

D Proofs for Section 4

We first provide a lemma showing that a model can have accurate probability estimation of trajectories when it estimates the historical data well. The proof can be found in Appendix B of Liu et al. (2022a).

Lemma 9 (Restatement of Proposition 14 of Liu et al. (2022a))**.** *There exists a universal constant c such that for any $\delta \in (0, 1]$, with probability at least $1 - \delta$, for all $M \in \mathcal{M}$, it holds that*

$$\left(\sum_{\tau \in \prod_{h=1}^{H}(\mathcal{O}_h \times \mathcal{A})} \left|\mathbb{P}^{\pi_b}_{M}(\tau) - \mathbb{P}^{\pi_b}_{M^\star}(\tau)\right|\right)^2 \leq \frac{c\left(\sum_{i=1}^{n} \log \frac{\mathbb{P}^{\pi_b}_{M^\star}(\tau^{(i)})}{\mathbb{P}^{\pi_b}_{M}(\tau^{(i)})} + \log \frac{|\mathcal{M}|}{\delta}\right)}{n}$$

Since the ground-truth $M^\star \in \mathcal{M}$, for the MLE $\widehat{M}$, we further have

$$\sum_{\tau \in \prod_{h=1}^{H}(\mathcal{O}_h \times \mathcal{A})} \left|\mathbb{P}^{\pi_b}_{\widehat{M}}(\tau) - \mathbb{P}^{\pi_b}_{M^\star}(\tau)\right| \leq \sqrt{\frac{c \log \frac{|\mathcal{M}|}{\delta}}{n}} := \varepsilon_{\mathrm{MLE}}.$$

In addition, marginalizing two distributions will not increase the TV distance, for any $h \in [H-1]$, the following inequalities hold,

$$\sum_{\tau_h, o_{h+1}} \left|\mathbb{P}^{\pi_b}_{\widehat{M}}(\tau_h, o_{h+1}) - \mathbb{P}^{\pi_b}_{M^*}(\tau_h, o_{h+1})\right| \leq \varepsilon_{\mathrm{MLE}}, \tag{3}$$

$$\sum_{\tau_h, f_{h+1}} \left|\mathbb{P}^{\pi_b}_{\widehat{M}}(\tau_h, f_{h+1}) - \mathbb{P}^{\pi_b}_{M^*}(\tau_h, f_{h+1})\right| \leq \varepsilon_{\mathrm{MLE}}. \tag{4}$$

D.1 Proof of Theorem 4

To begin with, we first state the following lemma for our operators $\mathbf{B}$, which is adapted from Lemma 32 of Liu et al. (2022a). We use $\mathbf{B}_{h:j+1}$ to denote $\prod_{h'=j+1}^{h} \mathbf{B}_{h'}(o_{h'}, a_{h'})$.

Lemma 10. *For any index $0 \le j < h \le H-1$, trajectory $\tau_j \in \prod_{h'=1}^{j}(\mathcal{O}_{h'} \times \mathcal{A})$, vector $x \in \mathbb{R}^O$, policy π, and operator $\mathbf{B}$ satisfying Assumption E, we have*

$$\sum_{\tau_{h:j+1}} \|\mathbf{B}_{h:j+1} x\|_1 \times \pi(\tau_{h:j+1} \mid \tau_j) \le C_{\mathcal{O}} \|x\|_1.$$

Proof. From the definition of $\mathbf{B}_{h:j+1}$, we have

$$\mathbf{B}_{h:j+1} x = \mathbf{B}_{h:j+1} \mathbb{O}_{j+1} \mathbb{O}_{j+1}^{\dagger,w} x.$$

Then, for any standard basis $\mathbf{e}_i \in \mathbb{R}^S$, we obtain

$$\begin{aligned}
&\sum_{\tau_{h:j+1}} \|\mathbf{B}_{h:j+1} \mathbb{O}_{j+1} \mathbf{e}_i\|_1 \times \pi(\tau_{h:j+1} \mid \tau_j) \\
=&\sum_{o} \sum_{\tau_{h:j+1}} \mathbb{P}_M(o_{h+1} = o, \tau_{h:j+1} \mid s_{j+1} = i) \cdot \pi(\tau_{h:j+1} \mid \tau_j) \\
=&\sum_{o} \sum_{\tau_{h:j+1}} \mathbb{P}_M^{\pi}(o_{h+1} = o, \tau_{h:j+1} \mid \tau_j, s_{j+1} = i) = 1.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sum_{\tau_{h:j+1}} \|\mathbf{B}_{h:j+1} x\|_1 \times \pi(\tau_{h:j+1} \mid \tau_j) \\
\le& \sum_{\tau_{h:j+1}} \|\mathbf{B}_{h:j+1} \mathbb{O}_{j+1}\|_1 \|\mathbb{O}_{j+1}^{\dagger,w} x\|_1 \times \pi(\tau_{h:j+1} \mid \tau_j) \\
\le& \|\mathbb{O}_{j+1}^{\dagger,w} x\|_1 \le C_{\mathcal{O}} \|x\|_1.
\end{aligned}$$

The last step is from Assumption E and $\|\mathbb{O}_{j+1}^{\dagger,w}\|_1 \le \|\Sigma_{\mathcal{O},h}^{-1}\|_1 \|\mathbb{O}_h^{\top} W_h^{-1}\|_1 \le C_{\mathcal{O}}$. □

Then, we prove Theorem 4. When the context is clear, we omit the (o_h, a_h) in $\mathbf{B}_h(o_h, a_h)$.

Proof. We first bound the single-step estimation error of π_b. For any $h \in [H-1]$, we have

$$\begin{aligned}
&\sum_{\tau_h} \pi_b(\tau_h) \times \left\| \left(\widehat{\mathbf{B}}_h(o_h, a_h) - \mathbf{B}_h(o_h, a_h) \right) \mathbf{b}(\tau_{h-1}) \right\|_1 \\
\le& \sum_{\tau_h} \pi_b(\tau_h) \times \left\| \widehat{\mathbf{B}}_h(o_h, a_h) \widehat{\mathbf{b}}(\tau_{h-1}) - \mathbf{B}_h(o_h, a_h) \mathbf{b}(\tau_{h-1}) \right\|_1 \\
+& \sum_{\tau_h} \pi_b(\tau_h) \times \left\| \widehat{\mathbf{B}}_h(o_h, a_h) \left(\widehat{\mathbf{b}}(\tau_{h-1}) - \mathbf{b}(\tau_{h-1}) \right) \right\|_1 \\
\le& 2 C_{\mathcal{O}} \varepsilon_{\text{MLE}}. \qquad (5)
\end{aligned}$$

For the last step, the first term is bounded by ε_{MLE} due to Eq. (3). The second term is from

$$\begin{aligned}
&\sum_{\tau_h} \pi_b(\tau_h) \times \left\| \widehat{\mathbf{B}}_h(o_h, a_h) \left(\widehat{\mathbf{b}}(\tau_{h-1}) - \mathbf{b}(\tau_{h-1}) \right) \right\|_1 \\
=& \sum_{\tau_{h-1}} \pi_b(\tau_{h-1}) \sum_{o_h, a_h} \pi_b(a_h \mid \tau_{h-1}, o_h) \left\| \widehat{\mathbf{B}}_h(o_h, a_h) \left(\widehat{\mathbf{b}}(\tau_{h-1}) - \mathbf{b}(\tau_{h-1}) \right) \right\|_1 \\
\le& C_{\mathcal{O}} \sum_{\tau_{h-1}} \pi_b(\tau_{h-1}) \left\| \widehat{\mathbf{b}}(\tau_{h-1}) - \mathbf{b}(\tau_{h-1}) \right\|_1 \qquad \text{(Lemma 10)} \\
\le& C_{\mathcal{O}} \varepsilon_{\text{MLE}}. \qquad \text{(Eq. (3))}
\end{aligned}$$

Then we consider the evaluation error of π_e. We decompose the evaluation error into single-step error

$$\begin{aligned}
\left|J(\pi_e) - J_{\widehat{M}}(\pi_e)\right| &= \sum_{h=1}^{H} \left| \sum_{o_h} (\mathbb{P}^{\pi_e}_{\widehat{M}}(o_h) - \mathbb{P}^{\pi_e}_{M^*}(o_h)) r(o_h) \right| \\
&\le \sum_{h=1}^{H} \sum_{o_h} \left| \mathbb{P}^{\pi_e}_{\widehat{M}}(o_h) - \mathbb{P}^{\pi_e}_{M^*}(o_h) \right| \\
&= \sum_{h=1}^{H} \sum_{o_h} \left| \sum_{\tau_{h-1}} (\mathbb{P}^{\pi_e}_{\widehat{M}}(o_h, \tau_{h-1}) - \mathbb{P}^{\pi_e}_{M^*}(o_h, \tau_{h-1})) \right| \\
&= \sum_{h=1}^{H} \sum_{o_h} \left| \mathbf{e}_{o_h}^{\top} \left(\sum_{\tau_{h-1}} \left(\widehat{\mathbf{B}}_{h-1} \cdots \widehat{\mathbf{B}}_1 \widehat{\mathbf{b}}_0 - \mathbf{B}_{h-1} \cdots \mathbf{B}_1 \mathbf{b}_0 \right) \times \pi_e(\tau_{h-1}) \right) \right| \\
&= \sum_{h=1}^{H} \left\| \sum_{\tau_{h-1}} \left(\widehat{\mathbf{B}}_{h-1} \cdots \widehat{\mathbf{B}}_1 \widehat{\mathbf{b}}_0 - \mathbf{B}_{h-1} \cdots \mathbf{B}_1 \mathbf{b}_0 \right) \times \pi_e(\tau_{h-1}) \right\|_1 .
\end{aligned}$$

For the case $h = 1$, according to Eq. (3), we have $\sum_{o_1} \left| \mathbb{P}^{\pi_e}_{\widehat{M}}(o_1) - \mathbb{P}^{\pi_e}_{M^*}(o_1) \right| \le \varepsilon_{\text{MLE}}$. For $2 \le h \le H$, we use the telescoping and obtain the following equality for any τ_{h-1}.

$$\widehat{\mathbf{B}}_{h-1} \cdots \widehat{\mathbf{B}}_1 \widehat{\mathbf{b}}_0 - \mathbf{B}_{h-1} \cdots \mathbf{B}_1 \mathbf{b}_0 = \sum_{j=1}^{h-1} \widehat{\mathbf{B}}_{h-1:j+1} (\widehat{\mathbf{B}}_j - \mathbf{B}_j) \mathbf{b}(\tau_{j-1}) + \widehat{\mathbf{B}}_{h-1:1} (\widehat{\mathbf{b}}_0 - \mathbf{b}_0). \quad (6)$$

Therefore, for a fixed step $2 \le h \le H$, we have

$$\begin{aligned}
&\left\| \sum_{\tau_{h-1}} \left(\widehat{\mathbf{B}}_{h-1} \cdots \widehat{\mathbf{B}}_1 \widehat{\mathbf{b}}_0 - \mathbf{B}_{h-1} \cdots \mathbf{B}_1 \mathbf{b}_0 \right) \times \pi_e(\tau_{h-1}) \right\|_1 . \\
\le &\left\| \sum_{\tau_{h-1}} \pi_e(\tau_{h-1}) \widehat{\mathbf{B}}_{h-1:1} (\widehat{\mathbf{b}}_0 - \mathbf{b}_0) \right\|_1 + \left\| \sum_{\tau_{h-1}} \pi_e(\tau_{h-1}) \sum_{j=1}^{h-1} \widehat{\mathbf{B}}_{h-1:j+1} (\widehat{\mathbf{B}}_j - \mathbf{B}_j) \mathbf{b}(\tau_{j-1}) \right\|_1 .
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
\left\| \sum_{\tau_{h-1}} \pi_e(\tau_{h-1}) \widehat{\mathbf{B}}_{h-1:1} (\widehat{\mathbf{b}}_0 - \mathbf{b}_0) \right\|_1 &\le \sum_{\tau_{h-1}} \left\| \widehat{\mathbf{B}}_{h-1:1} (\widehat{\mathbf{b}}_0 - \mathbf{b}_0) \right\|_1 \times \pi_e(\tau_{h-1}) \\
&\le C_{\mathcal{O}} \| \widehat{\mathbf{b}}_0 - \mathbf{b}_0 \|_1 && \text{(Lemma 10)} \\
&\le C_{\mathcal{O}} \varepsilon_{\text{MLE}}. && \text{(Eq. (3))}
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
&\left\| \sum_{\tau_{h-1}} \pi_e(\tau_{h-1}) \sum_{j=1}^{h-1} \widehat{\mathbf{B}}_{h-1:j+1} (\widehat{\mathbf{B}}_j - \mathbf{B}_j) \mathbf{b}(\tau_{j-1}) \right\|_1 \\
= &\left\| \sum_{j=1}^{h-1} \sum_{\tau_j} \pi_e(\tau_j) \sum_{\tau_{h-1:j+1}} \pi_e(\tau_{h-1:j+1} \mid \tau_j) \times \widehat{\mathbf{B}}_{h-1:j+1} (\widehat{\mathbf{B}}_j - \mathbf{B}_j) \mathbf{b}(\tau_{j-1}) \right\|_1 \\
= &\left\| \sum_{j=1}^{h-1} \sum_{\tau_j} \pi_e(\tau_j) \mathbf{P}_{\tau_j} \widehat{\mathbb{O}}^{\dagger,w}_{j+1} (\widehat{\mathbf{B}}_j - \mathbf{B}_j) \mathbf{b}(\tau_{j-1}) \right\|_1 , \qquad (7)
\end{aligned}$$

where

$$\mathbf{P}_{\tau_j} = \sum_{\tau_{h-1:j+1}} \pi_e(\tau_{h-1:j+1} \mid \tau_j) \times \widehat{\mathbf{B}}_{h-1:j+1} \widehat{\mathbb{O}}_{j+1}$$

and we observe that

$$\begin{aligned}[\mathbf{P}_{\tau_j}]_{ik} &= \sum_{\tau_{h-1:j+1}} \pi_e(\tau_{h-1:j+1} \mid \tau_j)\mathbb{P}_{\widehat{M}}(o_h = i, o_{h-1:j+1} \mid s_{j+1} = k, a_{h:j+1}) \\ &= \mathbb{P}^{\pi_e}_{\widehat{M}}(o_h = i \mid s_{j+1} = k, \tau_j).\end{aligned}$$

Therefore $\|\mathbf{P}_{\tau_j}\|_1 = 1$. Since $\|\widehat{\mathbb{O}}^{\dagger,w}_{j+1}\|_1 \le C_{\mathcal{O}}$, we further upper bound Eq. (7) as

$$\begin{aligned}&\left\|\sum_{j=1}^{h-1}\sum_{\tau_j}\pi_e(\tau_j)\mathbf{P}_{\tau_j}\widehat{\mathbb{O}}^{\dagger,w}_{j+1}(\widehat{\mathbf{B}}_j - \mathbf{B}_j)\mathbf{b}(\tau_{j-1})\right\|_1 \\ &\le C_{\mathcal{O}}\sum_{j=1}^{h-1}\sum_{\tau_j}\pi_e(\tau_j)\left\|(\widehat{\mathbf{B}}_j - \mathbf{B}_j)\mathbf{b}(\tau_{j-1})\right\|_1.\end{aligned}$$

Taking summation over h from 1 to H, we obtain

$$\left|J(\pi_e) - J_{\widehat{M}}(\pi_e)\right| \le HC_{\mathcal{O}}\varepsilon_{\text{MLE}} + HC_{\mathcal{O}}\sum_{j=1}^{H-1}\sum_{\tau_j}\pi_e(\tau_j)\left\|(\widehat{\mathbf{B}}_j - \mathbf{B}_j)\mathbf{b}(\tau_{j-1})\right\|_1. \tag{8}$$

From Eq. (5), we know that $\sum_{\tau_h}\pi_b(\tau_h)\times\left\|\left(\widehat{\mathbf{B}}_h - \mathbf{B}_h\right)\mathbf{b}(\tau_{h-1})\right\|_1 \le 2C_{\mathcal{O}}\varepsilon_{\text{MLE}}$. Therefore, we consider the ratio between estimation error of π_e and estimation error of π_b for step $h \in [H-1]$.

$$\begin{aligned}&\frac{\sum_{\tau_h}\pi_e(\tau_h)\times\left\|(\widehat{\mathbf{B}}_h - \mathbf{B}_h)\mathbf{b}(\tau_{h-1})\right\|_1}{\sum_{\tau_h}\pi_b(\tau_h)\times\left\|\left(\widehat{\mathbf{B}}_h - \mathbf{B}_h\right)\mathbf{b}(\tau_{h-1})\right\|_1} \\ &= \frac{\sum_{\tau_h}\pi_e(\tau_h)\mathbb{P}_{M^*}(\tau_{h-1})\times\left\|(\widehat{\mathbf{B}}_h - \mathbf{B}_h)\mathbb{O}_h\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\right\|_1}{\sum_{\tau_h}\pi_b(\tau_h)\mathbb{P}_{M^*}(\tau_{h-1})\times\left\|(\widehat{\mathbf{B}}_h - \mathbf{B}_h)\mathbb{O}_h\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\right\|_1} \\ &= \frac{\sum_{o_h,a_h}\mathbb{E}_{\pi_e}\left[\pi_e(a_h \mid \tau_{h-1}, o_h)\|(\widehat{\mathbf{B}}_h - \mathbf{B}_h)\mathbb{O}_h\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_1\right]}{\sum_{o_h,a_h}\mathbb{E}_{\pi_b}\left[\pi_b(a_h \mid \tau_{h-1}, o_h)\|(\widehat{\mathbf{B}}_h - \mathbf{B}_h)\mathbb{O}_h\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_1\right]}.\end{aligned}$$

In the first step, we use the relation $\mathbf{b}(\tau_h) = \mathbb{P}_{M^\star}(\tau_h)\mathbb{O}_{h+1}\mathbf{b}_{\mathcal{S}}(\tau_h)$. Combining it with Eq. (8), we have

$$\left|J(\pi_e) - J_{\widehat{M}}(\pi_e)\right| \le HC_{\mathcal{O}}\varepsilon_{\text{MLE}} + 2H^2C_{\mathcal{O}}^2C_{\text{eff},1}\varepsilon_{\text{MLE}},$$

where

$$C_{\text{eff},1} := \max_{h\in[H-1]}\frac{\sum_{o_h,a_h}\mathbb{E}_{\pi_e}\left[\pi_e(a_h \mid \tau_{h-1}, o_h)\|(\widehat{\mathbf{B}}_h - \mathbf{B}_h)\mathbb{O}_h\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_1\right]}{\sum_{o_h,a_h}\mathbb{E}_{\pi_b}\left[\pi_b(a_h \mid \tau_{h-1}, o_h)\|(\widehat{\mathbf{B}}_h - \mathbf{B}_h)\mathbb{O}_h\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_1\right]}.$$

We have proved the first part of the theorem. Next, we need to upper bound $C_{\text{eff},1}$. Let $x_l(o_h, a_h) \in \mathbb{R}^{|\mathcal{S}_h|}$ be the l-th row of $\left[(\widehat{\mathbf{B}}_h(o_h, a_h) - \mathbf{B}_h(o_h, a_h)\mathbb{O}_h\right]$, then we have

$$\begin{aligned}C_{\text{eff},1} &\le \max_{h\in[H-1]}\max_{o_h,a_h,l}\frac{\mathbb{E}_{\pi_e}\left[\pi_e(a_h \mid \tau_{h-1}, o_h)|x_l(o_h,a_h)^\top\mathbf{b}_{\mathcal{S}}(\tau_{h-1})|\right]}{\mathbb{E}_{\pi_b}\left[\pi_b(a_h \mid \tau_{h-1}, o_h)|x_l(o_h,a_h)^\top\mathbf{b}_{\mathcal{S}}(\tau_{h-1})|\right]} \\ &\le \max_{h\in[H-1]}\max_{o_h,a_h,l}C_{\mathcal{A}}\frac{\mathbb{E}_{\pi_e}|x_l(o_h,a_h)^\top\mathbf{b}_{\mathcal{S}}(\tau_{h-1})|}{\mathbb{E}_{\pi_b}|x_l(o_h,a_h)^\top\mathbf{b}_{\mathcal{S}}(\tau_{h-1})|}.\end{aligned}$$

To lower bound the denominator, we have

$$\mathbb{E}_{\pi_b}\left|x_l(o_h,a_h)^\top\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\right| = \mathbb{E}_{\pi_b}\frac{\left(x_l(o_h,a_h)^\top\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\right)^2}{|x_l(o_h,a_h)^\top\mathbf{b}_{\mathcal{S}}(\tau_{h-1})|}$$

$$
\begin{aligned}
&\geq \mathbb{E}_{\pi_b} \frac{\big(x_l(o_h,a_h)^\top \mathbf{b}_{\mathcal{S}}(\tau_{h-1})\big)^2}{\|x_l(o_h,a_h)\|_2 \|\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_2}. \\
&\geq \frac{\sigma_{\min}(\Sigma_{\mathcal{H},h})\|x_l(o_h,a_h)\|_2^2}{\|x_l(o_h,a_h)\|_2\|\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_2} \\
&\geq \sigma_{\min}(\Sigma_{\mathcal{H},h})\|x_l(o_h,a_h)\|_2. \qquad (9)
\end{aligned}
$$

Here $\Sigma_{\mathcal{H},h} = \mathbb{E}_{\pi_b}\left[\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\mathbf{b}_{\mathcal{S}}(\tau_{h-1})^\top\right]$. The first inequality is from Cauchy-Schwarz inequality and the last inequality is because $\|\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_2 \leq 1$. The numerator is upper bounded by

$$
\left|\mathbb{E}_{\pi_e}\left[x_l(o_h,a_h)^\top \mathbf{b}_{\mathcal{S}}(\tau_{h-1})\right]\right| \leq \|x_l(o_h,a_h)\|_2 \|\mathbb{E}_{\pi_e}\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_2. \qquad (10)
$$

Combining Eq. (9) and Eq. (10),

$$
C_{\text{eff},1} \leq C_{\mathcal{A}} C_{\mathcal{H}}.
$$

The proof is completed. □

D.2 Proof of Theorem 5

As in Appendix D.1, we first show a lemma for the operators $\mathbf{B}$ in the multi-step outcome scenario.

Lemma 11. *For any index $0 \leq j < h \leq H-1$, trajectory $\tau_j \in \prod_{h'=1}^{j}(\mathcal{O}_{h'} \times \mathcal{A})$, vector $x \in \mathbb{R}^{|\mathcal{F}_{j+1}|}$, policy π, and operator $\mathbf{B}$ satisfying Assumption D, we have*

$$
\sum_{\tau_{h:j+1}} \|\mathbf{B}_{h:j+1} x\|_1 \times \pi(\tau_{h:j+1} \mid \tau_j) \leq C_{\mathcal{F}} \|x\|_1.
$$

Proof. From the definition of $\mathbf{B}_{h:j+1}$, we have

$$
\mathbf{B}_{h:j+1} x = \mathbf{B}_{h:j+1} U_{\mathcal{F},j+1} U_{\mathcal{F},j+1}^{\dagger,w} x.
$$

Then, for any standard basis $\mathbf{e}_i \in \mathbb{R}^S$, we obtain

$$
\begin{aligned}
&\sum_{\tau_{h:j+1}} \|\mathbf{B}_{h:j+1} U_{\mathcal{F},j+1} \mathbf{e}_i\|_1 \times \pi(\tau_{h:j+1} \mid \tau_j) \\
&= \sum_{\tau_{h:j+1}} \sum_{f_{h+1}} \mathbb{P}_M^{\pi_b}(f_{h+1}, o_{h:j+1} \mid s_{j+1} = i, a_{h:j+1}) \cdot \pi(\tau_{h:j+1} \mid \tau_j) \\
&= \sum_{\tau_{h:j+1}} \sum_{f_{h+1}} \mathbb{P}_M^{a_{h:j+1} \sim \pi, a_{h+1:} \sim \pi_b}(f_{h+1}, \tau_{h:j+1} \mid \tau_j, s_{j+1} = i) = 1.
\end{aligned}
$$

Therefore

$$
\begin{aligned}
&\sum_{\tau_{h:j+1}} \|\mathbf{B}_{h:j+1} x\|_1 \times \pi(\tau_{h:j+1} \mid \tau_j) \\
&\leq \sum_{\tau_{h:j+1}} \|\mathbf{B}_{h:j+1} U_{\mathcal{F},j+1}\|_1 \|U_{\mathcal{F},j+1}^{\dagger,w} x\|_1 \times \pi(\tau_{h:j+1} \mid \tau_j) \\
&\leq C_{\mathcal{F}} \|x\|_1.
\end{aligned}
$$

The last step is from Assumption D and $\|U_{\mathcal{F},j+1}^{\dagger,w}\|_1 \leq \|\Sigma_{\mathcal{F},j+1}^{-1}\|_1 \|U_{\mathcal{F},j+1}^\top Z_{j+1}^{-1}\|_1 = C_{\mathcal{F}}$. □

Then, we prove Theorem 5.

Proof. We first observe that for any $h \in [H-1]$,

$$
\sum_{\tau_h} \pi_b(\tau_h) \times \|\widehat{\mathbf{b}}(\tau_h) - \mathbf{b}(\tau_h)\|_1 = \sum_{\tau_h, f_{h+1}} \left|\mathbb{P}_{\widehat{M}}^{\pi_b}(\tau_h, f_{h+1}) - \mathbb{P}_{M^*}^{\pi_b}(\tau_h, f_{h+1})\right| \leq \varepsilon_{\text{MLE}}.
$$

The inequality is from Eq. (4). Hence, using Lemma 11, for the estimation error of π_b at step $h \in [H-1]$, we have

$$\begin{aligned}
&\sum_{\tau_h} \pi_b(\tau_h) \times \left\| \left(\widehat{\mathbf{B}}_h(o_h, a_h) - \mathbf{B}_h(o_h, a_h)\right) \mathbf{b}(\tau_{h-1}) \right\|_1 \\
\leq &\sum_{\tau_h} \pi_b(\tau_h) \times \left\| \widehat{\mathbf{B}}_h(o_h, a_h)\widehat{\mathbf{b}}(\tau_{h-1}) - \mathbf{B}_h(o_h, a_h)\mathbf{b}(\tau_{h-1}) \right\|_1 \\
+ &\sum_{\tau_h} \pi_b(\tau_h) \times \left\| \widehat{\mathbf{B}}_h(o_h, a_h) \left(\widehat{\mathbf{b}}(\tau_{h-1}) - \mathbf{b}(\tau_{h-1})\right) \right\|_1 \\
\leq &2C_{\mathcal{F}}\varepsilon_{\text{MLE}}. \qquad (11)
\end{aligned}$$

Next, we consider the reward of π_e at each step. For step h, we construct policy μ_h, which follows π_e until step $h-1$ and follows π_b from step h. It is clear that μ_h has the same expected reward as π_e at step h. Let $r_h(f_h)$ be the reward of f_h at step h, we have

$$\begin{aligned}
\left|J(\pi_e) - J_{\widehat{M}}(\pi_e)\right| &= \sum_{h=1}^{H} \left| \sum_{f_h} (\mathbb{P}^{\mu_h}_{\widehat{M}}(f_h) - \mathbb{P}^{\mu_h}_{M^*}(f_h)) r_h(f_h) \right| \\
&\leq \sum_{h=1}^{H} \sum_{f_h} \left| \mathbf{e}_{f_h}^\top \left(\sum_{\tau_{h-1}} \left(\widehat{\mathbf{B}}_{h-1} \cdots \widehat{\mathbf{B}}_1 \widehat{\mathbf{b}}_0 - \mathbf{B}_{h-1} \cdots \mathbf{B}_1 \mathbf{b}_0 \right) \times \pi_e(\tau_{h-1}) \right) \right| \\
&= \sum_{h=1}^{H} \left\| \sum_{\tau_{h-1}} \left(\widehat{\mathbf{B}}_{h-1} \cdots \widehat{\mathbf{B}}_1 \widehat{\mathbf{b}}_0 - \mathbf{B}_{h-1} \cdots \mathbf{B}_1 \mathbf{b}_0 \right) \times \pi_e(\tau_{h-1}) \right\|_1 .
\end{aligned}$$

Similar to the derivation of Theorem 4, for the case $h = 1$, according to Eq. (4), we have $\sum_{f_1} \left| \mathbb{P}^{\pi_b}_{\widehat{M}}(f_1) - \mathbb{P}^{\pi_b}_{M^*}(f_1) \right| \leq \varepsilon_{\text{MLE}}$. For $2 \leq h \leq H$, we use the telescoping in Eq. (6) and obtain

$$\begin{aligned}
&\left\| \sum_{\tau_{h-1}} \left(\widehat{\mathbf{B}}_{h-1} \cdots \widehat{\mathbf{B}}_1 \widehat{\mathbf{b}}_0 - \mathbf{B}_{h-1} \cdots \mathbf{B}_1 \mathbf{b}_0 \right) \times \pi_e(\tau_{h-1}) \right\|_1 . \\
\leq &\left\| \sum_{\tau_{h-1}} \pi_e(\tau_{h-1}) \widehat{\mathbf{B}}_{h-1:1} (\widehat{\mathbf{b}}_0 - \mathbf{b}_0) \right\|_1 + \left\| \sum_{\tau_{h-1}} \pi_e(\tau_{h-1}) \sum_{j=1}^{h-1} \widehat{\mathbf{B}}_{h-1:j+1} (\widehat{\mathbf{B}}_j - \mathbf{B}_j) \mathbf{b}(\tau_{j-1}) \right\|_1 .
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
\left\| \sum_{\tau_{h-1}} \pi_e(\tau_{h-1}) \widehat{\mathbf{B}}_{h-1:1} (\widehat{\mathbf{b}}_0 - \mathbf{b}_0) \right\|_1 &\leq \sum_{\tau_{h-1}} \left\| \widehat{\mathbf{B}}_{h-1:1} (\widehat{\mathbf{b}}_0 - \mathbf{b}_0) \right\|_1 \times \pi_e(\tau_{h-1}) \\
&\leq C_{\mathcal{F}} \|\widehat{\mathbf{b}}_0 - \mathbf{b}_0\|_1 \qquad \text{(Lemma 11)} \\
&\leq C_{\mathcal{F}} \varepsilon_{\text{MLE}}. \qquad \text{(Eq. (4))}
\end{aligned}$$

For the second term, we have

$$\begin{aligned}
&\left\| \sum_{\tau_{h-1}} \pi_e(\tau_{h-1}) \sum_{j=1}^{h-1} \widehat{\mathbf{B}}_{h-1:j+1} (\widehat{\mathbf{B}}_j - \mathbf{B}_j) \mathbf{b}(\tau_{j-1}) \right\|_1 \\
= &\left\| \sum_{j=1}^{h-1} \sum_{\tau_j} \pi_e(\tau_j) \sum_{\tau_{h-1:j+1}} \pi_e(\tau_{h-1:j+1} \mid \tau_j) \times \widehat{\mathbf{B}}_{h-1:j+1} (\widehat{\mathbf{B}}_j - \mathbf{B}_j) \mathbf{b}(\tau_{j-1}) \right\|_1 \\
= &\left\| \sum_{j=1}^{h-1} \sum_{\tau_j} \pi_e(\tau_j) \mathbf{P}_{\tau_j} \widehat{U}^{\dagger, w}_{\mathcal{F}, j+1} (\widehat{\mathbf{B}}_j - \mathbf{B}_j) \mathbf{b}(\tau_{j-1}) \right\|_1 , \qquad (12)
\end{aligned}$$

where

$$\mathbf{P}_{\tau_j} = \sum_{\tau_{h-1:j+1}} \pi_e(\tau_{h-1:j+1} \mid \tau_j) \times \widehat{\mathbf{B}}_{h-1:j+1} \widehat{U}_{\mathcal{F},j+1}$$

and we observe that

$$\begin{aligned}[\mathbf{P}_{\tau_j}]_{ik} &= \sum_{\tau_{h-1:j+1}} \pi_e(\tau_{h-1:j+1} \mid \tau_j) \mathbb{P}^{\pi_b}_{\widehat{M}}(f_h = i, o_{h-1:j+1} \mid s_{j+1} = k, a_{h:j+1}) \\ &= \mathbb{P}_{\widehat{M}}^{a_{h-1:j+1}\sim\pi_e, a_{h:}\sim\pi_b}(f_h = i \mid s_{j+1} = k, \tau_j).\end{aligned}$$

Therefore $\|\mathbf{P}_{\tau_j}\|_1 = 1$. Since $\|\widehat{U}^{\dagger,w}_{\mathcal{F},j+1}\|_1 \le C_{\mathcal{F}}$, we further upper bound Eq. (12) as

$$\begin{aligned}&\left\| \sum_{j=1}^{h-1} \sum_{\tau_j} \pi_e(\tau_j) \mathbf{P}_{\tau_j} \widehat{U}^{\dagger,w}_{\mathcal{F},j+1} (\widehat{\mathbf{B}}_j - \mathbf{B}_j) \mathbf{b}(\tau_{j-1}) \right\|_1 \\ &\le C_{\mathcal{F}} \sum_{j=1}^{h-1} \sum_{\tau_j} \pi_e(\tau_j) \left\| (\widehat{\mathbf{B}}_j - \mathbf{B}_j) \mathbf{b}(\tau_{j-1}) \right\|_1 .\end{aligned}$$

Taking summation over h from 1 to H, we get

$$\left| J(\pi_e) - J_{\widehat{M}}(\pi_e) \right| \le H C_{\mathcal{O}} \varepsilon_{\text{MLE}} + H C_{\mathcal{F}} \sum_{j=1}^{H-1} \sum_{\tau_j} \pi_e(\tau_j) \left\| (\widehat{\mathbf{B}}_j - \mathbf{B}_j) \mathbf{b}(\tau_{j-1}) \right\|_1 . \tag{13}$$

According to Eq. (11), we have $\sum_{\tau_h} \pi_b(\tau_h) \times \left\| \left(\widehat{\mathbf{B}}_h - \mathbf{B}_h \right) \mathbf{b}(\tau_{h-1}) \right\|_1 \le 2 C_{\mathcal{F}} \varepsilon_{\text{MLE}}$. Therefore, the ratio between estimation error of π_e and estimation error of π_b for step $h \in [H-1]$ is written as:

$$\begin{aligned}&\frac{\sum_{\tau_h} \pi_e(\tau_h) \times \left\| (\widehat{\mathbf{B}}_h - \mathbf{B}_h) \mathbf{b}(\tau_{h-1}) \right\|_1}{\sum_{\tau_h} \pi_b(\tau_h) \times \left\| \left(\widehat{\mathbf{B}}_h - \mathbf{B}_h \right) \mathbf{b}(\tau_{h-1}) \right\|_1} \\ &= \frac{\sum_{\tau_h} \pi_e(\tau_h) \mathbb{P}_{M^*}(\tau_{h-1}) \times \left\| (\widehat{\mathbf{B}}_h - \mathbf{B}_h) U_{\mathcal{F},h} \mathbf{b}_{\mathcal{S}}(\tau_{h-1}) \right\|_1}{\sum_{\tau_h} \pi_b(\tau_h) \mathbb{P}_{M^*}(\tau_{h-1}) \times \left\| (\widehat{\mathbf{B}}_h - \mathbf{B}_h) U_{\mathcal{F},h} \mathbf{b}_{\mathcal{S}}(\tau_{h-1}) \right\|_1} \\ &= \frac{\sum_{o_h,a_h} \mathbb{E}_{\pi_e} \left[\pi_e(a_h \mid \tau_{h-1}, o_h) \| (\widehat{\mathbf{B}}_h - \mathbf{B}_h) U_{\mathcal{F},h} \mathbf{b}_{\mathcal{S}}(\tau_{h-1}) \|_1 \right]}{\sum_{o_h,a_h} \mathbb{E}_{\pi_b} \left[\pi_b(a_h \mid \tau_{h-1}, o_h) \| (\widehat{\mathbf{B}}_h - \mathbf{B}_h) U_{\mathcal{F},h} \mathbf{b}_{\mathcal{S}}(\tau_{h-1}) \|_1 \right]}.\end{aligned}$$

In the first step, we use the relation $\mathbf{b}(\tau_h) = \mathbb{P}_{M^\star}(\tau_h) U_{\mathcal{F},h+1} \mathbf{b}_{\mathcal{S}}(\tau_h)$. Combining it with Eq. (13), we have

$$\left| J(\pi_e) - J_{\widehat{M}}(\pi_e) \right| \le H C_{\mathcal{F}} \varepsilon_{\text{MLE}} + 2 H^2 C_{\mathcal{F}}^2 C_{\text{eff},m} \varepsilon_{\text{MLE}},$$

where

$$C_{\text{eff},m} := \max_{h \in [H-1]} \frac{\sum_{o_h,a_h} \mathbb{E}_{\pi_e} \left[\pi_e(a_h \mid \tau_{h-1}, o_h) \| (\widehat{\mathbf{B}}_h - \mathbf{B}_h) U_{\mathcal{F},h} \mathbf{b}_{\mathcal{S}}(\tau_{h-1}) \|_1 \right]}{\sum_{o_h,a_h} \mathbb{E}_{\pi_b} \left[\pi_b(a_h \mid \tau_{h-1}, o_h) \| (\widehat{\mathbf{B}}_h - \mathbf{B}_h) U_{\mathcal{F},h} \mathbf{b}_{\mathcal{S}}(\tau_{h-1}) \|_1 \right]}.$$

Similar to Theorem 4, we further have $C_{\text{eff},m} \le C_{\mathcal{A}} C_{\mathcal{H}}$, the proof is completed. □

E Proofs for Section 5

Proof of Theorem 8. Recall that

$$\varepsilon_{\text{approx}} := \min_{M \in \mathcal{M}} \frac{1}{n} \sum_{i=1}^{n} \left(\log \mathbb{P}^{\pi}_{M^\star}(\tau_H^{(i)}) - \log \mathbb{P}^{\pi}_{M}(\tau_H^{(i)}) \right).$$

By invoking Lemma 9, with probability at least $1-\delta$, we have

$$\sum_{\tau\in\prod_{h=1}^{H}(\mathcal{O}_h\times\mathcal{A})}\left|\mathbb{P}^{\pi_b}_{\widehat{M}}(\tau)-\mathbb{P}^{\pi_b}_{M^\star}(\tau)\right| \leq \mathcal{O}\left(\sqrt{\frac{\log\frac{|\mathcal{M}|}{\delta}}{n}}+\varepsilon_{\text{approx}}\right) := \widetilde{\varepsilon}_{\text{MLE}}.$$

Therefore, for any $h\in[H-1]$, the following inequality holds,

$$\sum_{\tau_h,f_{h+1}}\left|\mathbb{P}^{\pi_b}_{\widehat{M}}(\tau_h,f_{h+1})-\mathbb{P}^{\pi_b}_{M^*}(\tau_h,f_{h+1})\right|\leq\widetilde{\varepsilon}_{\text{MLE}}.$$

The remaining proof is similar to the proof of Theorem 5. Finally, we get

$$\left|J(\pi_e)-J_{\widehat{M}}(\pi_e)\right|\leq HC_{\mathcal{F}}\widetilde{\varepsilon}_{\text{MLE}}+2H^2C_{\mathcal{F}}^2C_{\text{eff},m}\widetilde{\varepsilon}_{\text{MLE}}.$$

Substituting $\widetilde{\varepsilon}_{\text{MLE}}$ finishes the proof. □

F TIGHER BOUND FOR MEMORYLESS π_e

The paper mostly focuses on evaluation history-dependent target policies π_e in the previous section. Here, we show that we can obtain a tighter coverage coefficient for memoryless π_e. We present the results for multi-step outcome revealing, with the single-step case being similar.

Theorem 12. *Under the same condition as Theorem 5, and assuming π_e is memoryless, with probability at least $1-\delta$, we have*

$$\left|J(\pi_e)-J_{\widehat{M}}(\pi_e)\right|\leq\mathcal{O}\left(H^2C_{\mathcal{F}}^2\widetilde{C}_{\textit{eff},m}\sqrt{\frac{\log\frac{|\mathcal{M}|}{\delta}}{n}}\right).$$

where

$$\widetilde{C}_{\textit{eff},m}:=\max_{h\in[H-1]}\frac{\sum_{o_h,a_h}\sum_{l=1}^{|\mathcal{F}_{h+1}|}\left|\mathbb{E}_{\pi_e}\left[\pi_e(a_h\mid\tau_{h-1},o_h)\left[(\widehat{\mathbf{B}}_h-\mathbf{B}_h)U_{\mathcal{F},h}\right]_{l,:}\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\right]\right|}{\sum_{o_h,a_h}\mathbb{E}_{\pi_b}\left[\pi_b(a_h\mid\tau_{h-1},o_h)\|(\widehat{\mathbf{B}}_h-\mathbf{B}_h)U_{\mathcal{F},h}\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_1\right]}.$$

Note that the numerator of coverage coefficient $C_{\text{eff},m}$ in Theorem 5 can be written as:

$$\sum_{o_h,a_h}\sum_{l=1}^{|\mathcal{F}_{h+1}|}\mathbb{E}_{\pi_e}\left|\pi_e(a_h\mid\tau_{h-1},o_h)\left[(\widehat{\mathbf{B}}_h-\mathbf{B}_h)U_{\mathcal{F},h}\right]_{l,:}\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\right|.$$

Compared to $C_{\text{eff},m}$, the numerator in $\widetilde{C}_{\text{eff},m}$ places the absolute operator outside of the expectation, resulting in a tighter coverage measurement. Mathematically similar tighter coverage coefficients have also been discovered in offline linear MDPs or offline MDPs with general function approximation. We refer readers to Jiang & Xie (2024) for more details.

Proof of Theorem 12. The proof is the same as for Theorem 5 up to bounding Eq. (12). For memoryless π_e, we observe that $\mathbf{P}_{\tau_j}$ is the same across different τ_j. Therefore, we upper bound Eq. (12) as

$$\begin{aligned}&\left\|\sum_{j=1}^{h-1}\sum_{\tau_j}\pi_e(\tau_j)\mathbf{P}_{\tau_j}\widehat{U}^{\dagger,w}_{\mathcal{F},j+1}(\widehat{\mathbf{B}}_j-\mathbf{B}_j)\mathbf{b}(\tau_{j-1})\right\|_1\\&\leq C_{\mathcal{F}}\sum_{j=1}^{h-1}\left\|\sum_{\tau_j}\pi_e(\tau_j)(\widehat{\mathbf{B}}_j-\mathbf{B}_j)\mathbf{b}(\tau_{j-1})\right\|_1.\end{aligned}$$

Taking summation over h from 1 to H, we get

$$\left|J(\pi_e)-J_{\widehat{M}}(\pi_e)\right|\leq HC_{\mathcal{F}}\varepsilon_{\text{MLE}}+HC_{\mathcal{F}}\sum_{j=1}^{H-1}\left\|\sum_{\tau_j}\pi_e(\tau_j)(\widehat{\mathbf{B}}_j-\mathbf{B}_j)\mathbf{b}(\tau_{j-1})\right\|_1. \tag{14}$$

As before, we consider the ratio between estimation error of π_e and estimation error of π_b for step $h \in [H-1]$:

$$\begin{aligned}
&\frac{\left\|\sum_{\tau_h} \pi_e(\tau_h) \times (\widehat{\mathbf{B}}_h - \mathbf{B}_h)\mathbf{b}(\tau_{h-1})\right\|_1}{\sum_{\tau_h} \pi_b(\tau_h) \times \left\|\left(\widehat{\mathbf{B}}_h - \mathbf{B}_h\right)\mathbf{b}(\tau_{h-1})\right\|_1} \\
&= \frac{\left\|\sum_{\tau_h} \pi_e(\tau_h)\mathbb{P}_{M^*}(\tau_{h-1}) \times (\widehat{\mathbf{B}}_h - \mathbf{B}_h)U_{\mathcal{F},h}\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\right\|_1}{\sum_{\tau_h} \pi_b(\tau_h)\mathbb{P}_{M^*}(\tau_{h-1}) \times \left\|(\widehat{\mathbf{B}}_h - \mathbf{B}_h)U_{\mathcal{F},h}\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\right\|_1} \\
&= \frac{\sum_{o_h,a_h}\sum_{l=1}^{|\mathcal{F}_{h+1}|}\left|\mathbb{E}_{\pi_e}\left[\pi_e(a_h \mid \tau_{h-1}, o_h)\left[(\widehat{\mathbf{B}}_h - \mathbf{B}_h)U_{\mathcal{F},h}\right]_{l,:}\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\right]\right|}{\sum_{o_h,a_h}\mathbb{E}_{\pi_b}\left[\pi_b(a_h \mid \tau_{h-1}, o_h)\|(\widehat{\mathbf{B}}_h - \mathbf{B}_h)U_{\mathcal{F},h}\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_1\right]}.
\end{aligned}$$

Combining it with Eq. (14), we have

$$\left|J(\pi_e) - J_{\widehat{M}}(\pi_e)\right| \leq HC_{\mathcal{F}}\varepsilon_{\text{MLE}} + 2H^2C_{\mathcal{F}}^2\widetilde{C}_{\text{eff},m}\varepsilon_{\text{MLE}},$$

where

$$\widetilde{C}_{\text{eff},m} := \max_{h\in[H-1]} \frac{\sum_{o_h,a_h}\sum_{l=1}^{|\mathcal{F}_{h+1}|}\left|\mathbb{E}_{\pi_e}\left[\pi_e(a_h \mid \tau_{h-1}, o_h)\left[(\widehat{\mathbf{B}}_h - \mathbf{B}_h)U_{\mathcal{F},h}\right]_{l,:}\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\right]\right|}{\sum_{o_h,a_h}\mathbb{E}_{\pi_b}\left[\pi_b(a_h \mid \tau_{h-1}, o_h)\|(\widehat{\mathbf{B}}_h - \mathbf{B}_h)U_{\mathcal{F},h}\mathbf{b}_{\mathcal{S}}(\tau_{h-1})\|_1\right]}.$$

The proof is completed. □