A Minimax Learning Approach to Off-Policy Evaluation in Confounded Partially Observable Markov Decision Processes

Chengchun Shi*[1], Masatoshi Uehara*[2], Jiawei Huang [3], and Nan Jiang [3]

c.shi7@lse.ac.uk, mu223@cornell.edu, jiaweih@illinois.edu, nanjiang@illinois.edu
[1]Department of Statistics, London School of Economics and Political Science
[2]Department of Computer Science, Cornell University
[3]Department of Computer Science, University of Illinois Urbana-Champaign

Abstract

We consider off-policy evaluation (OPE) in Partially Observable Markov Decision Processes (POMDPs), where the evaluation policy depends only on observable variables and the behavior policy depends on unobservable latent variables. Existing works either assume no unmeasured confounders, or focus on settings where both the observation and the state spaces are tabular. In this work, we first propose novel identification methods for OPE in POMDPs with latent confounders, by introducing bridge functions that link the target policy's value and the observed data distribution. We next propose minimax estimation methods for learning these bridge functions, and construct three estimators based on these estimated bridge functions, corresponding to a value function-based estimator, a marginalized importance sampling estimator, and a doubly-robust estimator. Our proposal permits general function approximation and is thus applicable to settings with continuous or large observation/state spaces. The nonasymptotic and asymptotic properties of the proposed estimators are investigated in detail.

1 Introduction

Reinforcement learning (RL) has been successfully applied in online settings (e.g., video games) where interaction with environments is easy and the data can be adaptively generated. However, online interaction is often costly and dangerous for a number of high-stake domains ranging from health science to social science and economics. Offline (batch) reinforcement learning is concerned with policy learning and evaluation with limited and pre-collected data in a sample efficient manner Levine et al. (2020). The focus of this paper is the off-policy evaluation (OPE) problem, which refers to the task of estimating the value of an evaluation policy offline using data collected under a different behavior policy. It is critical in sequential decision-making problems from healthcare, robotics, and education where new policies need to be evaluated offline before online validation.

There is a growing literature on OPE (see e.g., Precup, 2000; Jiang and Li, 2016; Thomas and Brunskill, 2016; Liu et al., 2018; Xie et al., 2019; Yin and Wang, 2020; Liao et al., 2020; Yang et al., 2020; Pananjady and Wainwright, 2021; Kuzborskij et al., 2021; Zhang et al., 2021; Shi et al., 2021). A common assumption made in the aforementioned works is that of no unmeasured confounders. Specifically, they assume the time-varying state variables are fully observed and no unmeasured variables exist that confound the observed actions. However, this assumption is not testable from the data. It is often violated in observational datasets generated from healthcare applications.

*Equal contribution

To allow unmeasured confounders to exist, we model the observed data by a confounded Partially Observable Markov Decision Process (POMDP). Under this framework, the behavior policy is allowed to depend on some unobserved state variables that confound the action-reward association. The goal of OPE in confounded POMDPs is to estimate the value of an evaluation policy, which is a function of observed variables, using the data generated by such a behavior policy. This is a highly challenging problem. Directly applying the importance sampling methods Precup (2000); Liu et al. (2018) or the value function-based method Munos and Szepesvári (2008) would yield a biased estimator, as we do not have access to the unobserved state variables. Tennenholtz et al. (2020); Nair and Jiang (2021) have made important progresses on this problem by outlining a consistent OPE estimator in tabular settings. However, they are not applicable to settings with continuous observation/state spaces.

In this paper, we study OPE in non-tabular confounded POMDPs. Our contributions are summarized as follows. First, we provide a novel identification method for OPE with latent confounders. We only require the existence of two bridge functions, corresponding to a weight bridge function and a value bridge function. These bridge functions are defined as solutions to some integral equations. They can be interpreted as projections of the marginalized importance sampling weight and value functions defined on the latent state space onto the observation space. They are not always uniquely defined, but we do not require the uniqueness assumption to achieve consistent estimation.

Second, we propose minimax learning methods to estimate the two bridge functions with function approximation. The proposed method allows us to model these bridge functions via certain highly flexible function approximators (e.g., neural networks) and is thus applicable to settings with continuous or large observation/state spaces. Based on the estimated bridge functions, we further propose three estimators for the target policy's value, corresponding to a value function-based estimator, a marginalized importance sampling (IS) estimator and a doubly-robust (DR) estimator.

Finally, we systematically study the nonasymptotic and asymptotic properties of the proposed estimators. Specifically, we first show the finite-sample rate of convergence under the realizability and Bellman closedness assumptions on value bridge function. Similar assumptions are imposed on the Q- or density ratio function in fully observable MDP settings Munos and Szepesvári (2008); Uehara et al. (2021). Second, when the realizability assumption holds for both bridge functions, we establish the finite-sample rate of convergence of the proposed estimators without assuming Bellman closedness. Finally, when the bridge functions are uniquely defined, we prove that the DR estimator is asymptotically normal and achieves the Cramér-Rao lower bound. We expect our findings to contribute to the theoretical understanding of offline RL (see e.g., Munos and Szepesvári, 2008; Chen and Jiang, 2019; Kallus and Uehara, 2019).

The organization of the paper is as follows. In Section 2, we introduce the (confounded) POMDP setting and formulate the OPE problem. In Section 3, we focus on the bandit setting to illustrate the main idea of our proposal. In Section 4, we present our identification results in the time-homogeneous POMDP setting and propose the OPE estimators. In addition, we establish the asymptotic and nonasymptotic properties of the proposed estimators. In Section 5, we consider the time-inhomogeneous POMDP setting and develop methodologies to evaluating Markovian (reactive) evaluation policies. In Section 6, we show how to extend our results to evaluate fully history-dependent policies. In Section 7, we study the empirical performance of our methods in non-tabular environments.

1.1 Related Works

We discuss some related works on OPE with unmeasured confounders and spectral learning in this section. To save space, some additional related works on minimax estimation are discussed in Appendix A.

OPE with Unmeasured Confounders To handle unmeasured confounders, existing OPE methods can be roughly divided into the following three categories. The first type of methods proposes to develop partial identification bounds for the policy value based on sensitivity analysis Kallus and Zhou (2020); Namkoong et al. (2020); Zhang

and Bareinboim (2021). These methods rely on certain assumptions that might be difficult to validate in practice. For instance, Kallus and Zhou (2020) imposes a memoryless unmeasured confounding assumption.

The second type of methods derive the value estimates by making use of some auxiliary variables in the observed data (Zhang and Bareinboim, 2016; Wang et al., 2020; Liao et al., 2021; Ying et al., 2021; Bennett et al., 2021; Shi et al., 2022). These methods are not directly applicable to the POMDP setting we consider. For instance, Zhang and Bareinboim (2016); Wang et al. (2020); Liao et al. (2021) propose to model the observed data via a confounded MDP. In these settings, the time-varying observations satisfy the Markov property. However, the Markov assumption is violated in POMDPs.

The third type of methods adopt the POMDP model to formulate the confounded OPE problem and develop value estimators in tabular settings (Tennenholtz et al., 2020; Nair and Jiang, 2021). However, as we have commented, these methods are ineffective in non-tabular settings with continuous or large observation/state spaces.

We remark sequential importance sampling is applicable for OPE in non-confounded POMDPs (Futoma et al., 2020; Hu and Wager, 2021). However, they are not in confounded POMDPs, which we focus on.

Negative Controls Our proposal is closely related to a line of research on developing causal inference methods to evaluate the average treatment effect (ATE) with double-negative control adjustment. Among those works, Miao et al. (2018) outlined a consistent estimator for ATE with a categorical confounder. More recent methods generalize their approach to the continuous confounder setting Deaner (2018); Cui et al. (2020); Singh (2020); Ghassami et al. (2021); Xu et al. (2021); Mastouri et al. (2021); Kallus et al. (2021). All the aforementioned methods focus on a contextual bandit setting (i.e., one-shot decision making) and are not directly applicable to POMDPs that involve sequential decision making.

Spectral learning in POMDPs Our proposal is closely related to a line of works on developing spectral learning methods in the POMDP literature (Song et al., 2010; Boots et al., 2011; Hsu et al., 2012; Anandkumar et al., 2014; Kulesza et al., 2015; Hefny et al., 2015; Jin et al., 2020). These methods are originally designed for learning system dynamics in the absence of unmeasured confounders. They would yield biased value estimators in the presence of unmeasured confounders.

2 Problem Formulation

We consider a discounted Partially Observable Markov Decision Process (POMDP) defined as

$$\langle \mathcal{S}, \mathcal{A}, \mathcal{O}, \{r_t\}_{t\geq 0}, \{P_t\}_{t\geq 0}, \{Z_t\}_{t\geq 0}, \gamma \rangle,$$

where $\mathcal{S}$ denotes the state space, $\mathcal{A}$ denotes the finite action space, $\mathcal{O}$ denotes the observation space, $P_t : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ denotes the state transition kernel at time t, $Z_t : \mathcal{S} \to \Delta(\mathcal{O})$ denotes the observation function that defines the conditional distribution of the observation given the state at time t, $r_t : \mathcal{S} \times \mathcal{A} \to [0, 1]$ denotes the reward function that depends on the state-action-next state triplet at time t, and $\gamma \in [0, 1)$ is a discount factor that balances the immediate and future rewards. P_t, r_t and Z_t are unknown to us and need to be inferred from the observed data.

The data generating process in POMDPs is depicted in Figure 1. Suppose the environment is in a given state $s \in \mathcal{S}$. The agent selects an action $a \in \mathcal{A}$. Then the system transits into a new state s' and gives an immediate reward $r(s, a)$ to the agent. While we cannot directly observe the latent state s, we have access to an observation $o \sim Z(\cdot|s)$. A policy is a set of time-dependent decision rules $\pi = \{\pi_t\}$ where each $\pi_t : \mathcal{S} \times \mathcal{O} \to \Delta(\mathcal{A})$ maps the state-observation pair to a probability distribution over the action space. For a given policy π, the data generating process can be summarized as $S_0 \sim \nu_0, O_0 \sim Z_0(\cdot|S_0), A_0 \sim \pi_0(\cdot|S_0, O_0), R_0 = r_0(S_0, A_0), S_1 \sim P_0(\cdot|S_0, A_0), \cdots$ where ν_0 denotes the initial state distribution. The observed data up to t is given by $\tau_t = (O_0, A_0, R_0, \cdots, O_t, A_t, R_t)$. We denote its distribution by P_π. The policy value $J(\pi)$ is defined as the expected cumulative reward $\mathbb{E}_\pi[\sum_{t=0}^{H} \gamma^t R_t]$ for some $0 < H \leq +\infty$, where the expectation is taken w.r.t. the distribution of trajectories induced by policy π.

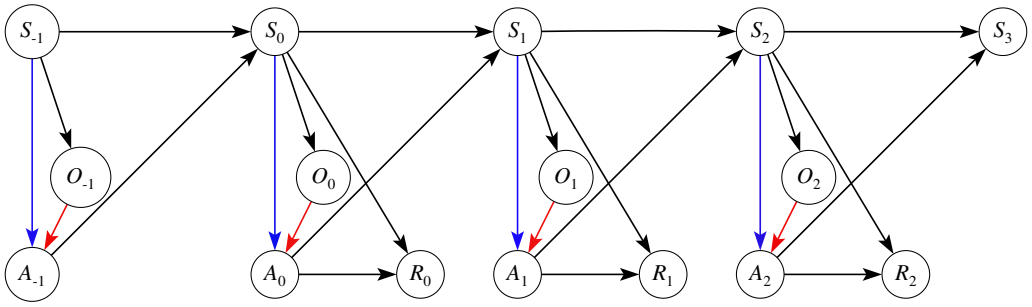


Figure 1: The data generating process in POMDPs. The evaluation policy in red depends on the observed variables. The behavior policy depends on the hidden state variables.

For a given evaluation policy $\pi^e = \{\pi_t^e\}_t$, the goal of OPE is to estimate $J(\pi^e)$ from an observational dataset that consists of N data trajectories, generated by a behavior policy $\pi^b = \{\pi_t^b\}_t$. Similar to Tennenholtz et al. (2020), we focus on the setting where π^b depends only on the latent state and π^e depends only on the observation. See also Figure 1. Under this model assumption, the state variables serve as a confounder between the action and the reward at each time. Since the state is not fully observed, the no unmeasured confounder assumption is violated.

3 Partially Observable Contextual Bandits

As a warm-up, we start with a partially observable contextual bandit setting (a special case of POMDP with horizon that equals 1) to illustrate the challenge of OPE with latent confounders and present our main idea. Suppose we have N data tuples that are i.i.d. $\{O_{-1}, A_0, O_0, R_0\}$ tuples where O_{-1} denotes the additional observation that is conditionally independent of (A_0, O_0, R_0) given S_0. The corresponding Bayesian network is depicted in Figure 2a.

If we were to observe S_0, following the standard OPE methods, we could identify the policy value using the IS or value function-based estimator, given by

$$\mathbb{E}\left[\eta_0(S_0, A_0)R_0\right] \text{ or } \mathbb{E}\left[v_0(S_0)\right] = \sum_a \mathbb{E}q_0(S_0, a)\pi_0^e(a \mid S_0),$$

where $q_0(S_0, A_0) = \mathbb{E}[R_0 \mid S_0, A_0]$, $\eta_0 = \pi_0^e/\pi_0^b$ denotes the IS ratio, and $\sum_a$ is an abbreviation of $\sum_{a\in\mathcal{A}}$. However, since we cannot observe S_0, these methods are not applicable. Naively replacing S_0 with O_0 would yield biased estimators.

To handle unmeasured confounders, we first assume the existence of certain bridge functions that link the target policy's value and the observed data distribution.

Assumption 1 (Existence of bridge functions)**.** *There exist functions $b'_V : \mathcal{A} \times \mathcal{O} \to \mathbb{R}, b'_W : \mathcal{A} \times \mathcal{O} \to \mathbb{R}$ that satisfy*

$$\mathbb{E}[b'_V(A_0, O_0) \mid S_0, A_0] = \mathbb{E}[R_0\pi^e(A_0 \mid O_0) \mid A_0, S_0], \tag{1}$$
$$\mathbb{E}[b'_W(A_0, O_{-1}) \mid S_0, A_0] = 1/\pi^b(A_0 \mid S_0). \tag{2}$$

We refer to $b'_V(\cdot)$ as a value bridge function and $b'_W(\cdot)$ as a weight bridge function.

By definition, weight bridge and value bridge functions can be interpreted as projections of the importance sampling weight (i.e., inverse propensity score) and value functions defined on the latent state space onto the observation space. When $\mathcal{S}$ and $\mathcal{O}$ are discrete, the existence of solutions to the integral equations in (1) and (2) is equivalent to certain matrix rank assumptions; see the assumptions in Nair and Jiang (2021); Boots et al.

(2011); Hsu et al. (2012). These assumptions require observed variables to contain sufficient information about unobserved states. Specifically, define a matrix $\Omega_a = \Pr_{\pi^b}(\mathbf{O}_0|A_0 = a, \mathbf{O}_{-1}))$ whose (i,j)-th element is given by $\Pr_{\pi^b}(\mathbf{O}_0 = o_i|A_0 = a, \mathbf{O}_{-1} = o_j))$ where o_i denotes the ith element in the observation space. Then they require $\mathrm{rank}(\Omega_a) = |\mathcal{S}|$ for any $a \in \mathcal{A}$. See Appendix B for details. When $\mathcal{S}$ and $\mathcal{O}$ are continuous, it follows from Picard's theorem (Carrasco et al., 2007, Theorem 2.41) that Assumption 1 is implied by certain completeness conditions. Finally, it is worthwhile to mention that we do *not* require these bridge functions to be *uniquely* defined.

Next, we show that Assumption 1 is a sufficient condition for policy value identification. The following theorem shows that if we know b'_V and b'_W in advance, we can consistently estimate the target value $J(\pi^e)$ from the offline data.

Lemma 1 (Pseudo identification formula)**.** *Suppose Assumption 1 holds. Then, for any b'_V, b'_W that solve (1),*

$$\begin{aligned} J(\pi^e) &= \mathbb{E}\left[\sum_a b'_V(a, O_0)\right] \quad \textit{and} \\ J(\pi^e) &= \mathbb{E}[b'_W(A_0, O_{-1}) R_0 \pi^e(A_0 \mid O_0)]. \end{aligned} \tag{3}$$

These formulas outline an IS estimator and a value function-based estimator for $J(\pi^e)$. It remains to identify the bridge functions b'_V and b'_W. However, even if Assumption 1 holds, it is still very challenging to learn b'_V and b'_W from the observed data as their definitions involve the unobserved state S_0. As such, we refer to Lemma 1 as the "pseudo" identification formula. In the following, we introduce some versions of bridge functions that are identifiable from the observed data.

Definition 1 (Learnable bridge functions)**.** *The learnable value bridge function $b_V : \mathcal{A} \times \mathcal{O} \to \mathbb{R}$ and learnable weight bridge function $b_W : \mathcal{A} \times \mathcal{O} \to \mathbb{R}$ are solutions to*

$$\begin{aligned} &\mathbb{E}[R_0 \pi^e(A_0|O_0)|A_0, O_{-1}] = \mathbb{E}[b_V(A_0, O_0)|A_0, O_{-1}] \\ &\quad \textit{and} \quad 1/P_{\pi^b}(A_0|O_0) = \mathbb{E}[b_W(A_0, O_{-1})|A_0, O_0]. \end{aligned} \tag{4}$$

Throughout this paper, we use b' to denote a bridge function and b to denote a *learnable* bridge function. The following lemma shows that any bridge function is a learnable bridge function.

Lemma 2. *Bridge functions defined as solutions to (1) and (2) are learnable bridge functions.*

We next present an equivalent definition for b_W. This characterization is useful when extending our results to the POMDP setting.

Lemma 3. $\mathbb{E}[b_W(A_0, O_{-1}) \mid A_0, O_0] = 1/\pi^b(A_0 \mid O_0)$ *holds if and only if*

$$\mathbb{E}[\sum_a f(O_0, a) - b_W(A_0, O_{-1}) f(O_0, A_0)] = 0, \quad \forall f.$$

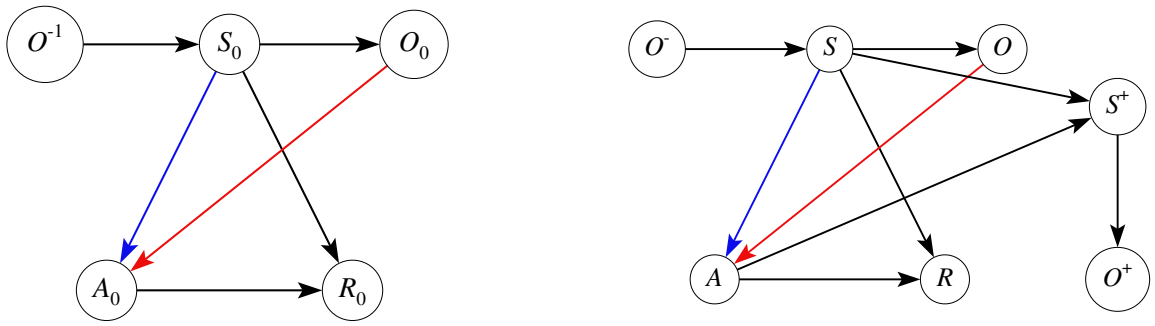
As we have commented, the class of bridge functions are difficult to estimate from the observed data. On the other hand, the class of learnable bridge functions is identifiable. However, they are *not* necessarily bridge functions. Thus, we cannot invoke Lemma 1 for value identification. Nonetheless, perhaps surprisingly, the following theorem shows that we can plug-in any learnable bridge function b_V and b_W for b'_V and b'_W in (3) under Assumption 1.

Theorem 1 (Key identification formula)**.** *Assume the existence of value bridge functions and learnable weight bridge functions. Then, letting $b_W(\cdot)$ be a learnable weight bridge function, we have*

$$J(\pi^e) = \mathbb{E}[b_W(A_0, O_{-1}) R_0].$$

Suppose the existence of weight bridge functions and learnable value bridge functions. Then, letting $b_V(\cdot)$ be a learnable value bridge function, we have

$$J(\pi^e) = \mathbb{E}[\sum_a b_V(a, O_0)].$$



(a) Partially Observable Contextual Bandits

(b) Time-homogeneous POMDPs.

Figure 2: Red lines depict the dependency of evaluation policies on the observed variables. Blue lines depict the dependence of behavior policies on the state variables.

We make a few remarks. First, the assumptions in the first and second statements do not imply each other. In both cases, Assumption 1, the existence of bridge functions, plays a critical role in ensuring the validity of the above equations. Specifically, when Assumption 1 holds, the assumptions in both the first and second statements are automatically satisfied, since bridge functions are learnable bridge functions. If we naively replace Assumption 1 with the existence of learnable bridge functions that satisfy equation 2, then Theorem 1 is no longer valid.

Second, Theorem 1 implies that we only need bridge functions to exist, but do not need to identify them. As long as they exist, learnable bridge functions can be used for estimation, which are not necessarily bridge functions.

Third, similar identification methods have been developed in the causal inference literature for evaluating the average treatment effect with double negative controls; see Appendix A. However, their settings differ from ours. For instance, those works require policies to be constant functions of negative controls. In contrast, we allow our evaluation policy to depend on the observation which serves as negative controls in their setting.

Finally, we show $J(\pi)$ can be uniquely identified by assuming the existence of value bridge functions only. Compared to Theorem 1, we do not require the existence of both the value and the (learnable) weight bridge functions. Instead, we impose the completeness assumption.

Theorem 2 (Key identification formula with completeness)**.** *Suppose learnable value bridge functions exist and the completeness assumption holds, i.e.,*

$$\mathbb{E}[g(S_0, A_0) \mid A_0, O_{-1}] = 0 \implies g(S_0, A_0) = 0.$$

Then, letting $b_V(\cdot)$ be a learnable value bridge function, we have

$$J(\pi^e) = \mathbb{E}[\sum_a b_V(a, O_0)].$$

We remark that in the tabular case, the completeness assumption is satisfied when $\Pr_{\pi_b}(\mathbf{S}_0 \mid \mathbf{O}_{-1}) = \text{rank}(|\mathcal{S}|)$. In addition, the value can be similarly identified when learnable weight bridge functions exist and the completeness condition is satisfied.

4 OPE in Time-homogeneous POMDPs

In this section, we focus on time-homogeneous POMDPs where $P_t, Z_t, r_t, \pi_t^e, \pi_t^b$ are stationary over time. We denote them by P, Z, r, π^e, π^b, respectively. Our goal is to estimate the target policy's value in the infinite horizon

setting ($H = \infty$). Consider a data tuple $(O^-, S, O, A, R, S^+, O^+)$ following

$$(O^-, S) \sim P_{\pi^b}(\cdot), A \sim \pi^b(\cdot|S), S^+ \sim P(\cdot|S, A),$$
$$R = r(S, A), O^+ \sim Z(\cdot|S^+),$$

where P_{π^b} denotes certain distribution over $\mathcal{O} \times \mathcal{S}$. The observed data tuple is given by (O^-, O, A, R, O^+). For example, for a data trajectory $\{(O_t, A_t, R_t)\}_{0 \le t \le n+1}$ under π_b, the observations can be summarized as $\{(O_{t-1}, O_t, A_t, R_t, O_{t+1})\}_{1 \le t \le n}$. The above configuration is satisfied with P_{π^b} equal to the occupancy distribution under π_b over $\mathcal{O} \times \mathcal{S}$. This conversion from trajectories to tuples is commonly used in the offline RL literature (Chen and Jiang, 2019). We assume the initial observation distribution $\nu_{\mathcal{O}}(o) \coloneqq \int Z(o \mid s)\nu(s)\mathrm{d}(s)$ is known to us. If unknown, it can be estimated by the empirical data distribution. With a slight abuse of notation, we denote the distribution of the data as $P_{\pi^b}(\cdot)$, e.g., $P_{\pi^b}(a \mid s) = \pi^b(a \mid s)$.

4.1 Identification

First, we extend our definitions of the bridge functions to the POMDP setting and require their existence in the assumption below. Let $w(s) = \sum_{t \ge 0} \gamma^t P_{\pi^e, t}(s)/P_{\pi^b}(s)$ denote the marginal density ratio where $P_{\pi^e, t}(\cdot)$ denotes the probability density function of S_t under the evaluation policy. In addition, let $q^{\pi^e}(a, s) = \mathbb{E}_{\pi^e}[\sum_{t \ge 0} \gamma^t R_t | A_0 = a, S_0 = s]$ denote the Q-function. These functions play an important role in constructing marginal IS and value function-based estimators in fully observable MDPs. The following bridge functions play similar roles in POMDPs.

Assumption 2 (Existence of bridge functions)**.** *There exist value bridge functions $b'_V : \mathcal{A} \times \mathcal{O} \to \mathbb{R}$ and weight bridge functions $b'_W : \mathcal{A} \times \mathcal{O} \to \mathbb{R}$ that satisfy*

$$\mathbb{E}[b'_V(A_0, O_0) \mid A_0, S_0] = \mathbb{E}_{\pi^e}[\sum_t \gamma^t R_t \pi^e(A_0 \mid O_0) | A_0, S_0],$$
$$\mathbb{E}[b'_W(A, O^-) \mid A, S] = \frac{w(S)}{P_{\pi^b}(A|S)}.$$

We make several remarks. First, the existence of b'_W implicitly requires a data coverage condition, i.e., for any s, $P_{\pi^b}(s) = 0 \Rightarrow \sum_{t \ge 0} \gamma^t p_t^{\pi^e}(s) = 0$. Second, when the states are fully observed and $\mathcal{S} = \mathcal{O}$, it follows immediately that $b'_V(a, s) = q^{\pi^e}(a, s)\pi^e(a \mid s)$ and $b'_W(a, s) = w(s)/\pi^b(a \mid s)$. Similar to the bandit setting, these bridge functions can thus be interpreted as projections of the value functions and the marginal IS weights onto the the observation space. Third, we do not require the bridge functions to be uniquely defined.

Next, we define the learnable bridge functions. They are consistent with those in Definition 1 when $\gamma = 0$.

Definition 2 (Learnable bridge functions)**.** *Learnable value bridge functions b_V are defined as solutions to*

$$\mathbb{E}[b_V(A, O) \mid A, O^-] = \mathbb{E}[R\pi^e(A \mid O) \mid A, O^-] + \gamma \mathbb{E}\left[\sum_{a'} b_V(a', O^+)\pi^e(A \mid O) \mid A, O^-\right]. \tag{5}$$

Learnable weight bridge functions b_W are defined as solutions to

$$\mathbb{E}[L_{\mathbb{W}}(b_W, f)] = 0, \quad \forall f : \mathcal{O} \times \mathcal{A} \to \mathbb{R} \tag{6}$$

where $L_{\mathbb{W}}(g, f)$ equals

$$\sum_{a'} \gamma g(A, O^-)\pi^e(A|O) f(O^+, a') - g(A, O^-) f(O, A) + (1 - \gamma)\mathbb{E}_{\tilde{O} \sim \nu_{\mathcal{O}}}[\sum_{a'} f(\tilde{O}, a')].$$

Finally, we show our key identification formula under Assumption 2. It extends Lemma 2 and Theorem 1 to the POMDP setting.

Theorem 3 (Key identification theorem)**.**

1. *Any bridge function is a learnable bridge function.*

2. *Suppose the existence of the value bridge function and learnable weight bridge function $b_W(\cdot)$. Then,*

$$J(\pi^e) = \frac{1}{1-\gamma}\mathbb{E}[b_W(A, O^-)R\pi(A \mid O)].$$

3. *Suppose the existence of the weight bridge function and learnable value bridge function $b_V(\cdot)$. Then,*

$$J(\pi^e) = \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_a b_V(a, \tilde{O})].$$

Similar to the bandit setting, any bridge function is a learnable bridge function, but the reverse is not true. However, the target policy's value can be consistently estimated based on any learnable bridge function. Theorem 3 forms the basis of our estimation procedure. It outlines a marginal IS estimator and a value function-based estimator for policy evaluation. conditions in 2 and 3 are automatically satisfied.

To identify policy values, we have so far imposed the following assumptions (a) the existence of learnable bridge functions and weight bridge functions; (b) the existence of learnable weight functions and value bridge functions. In the following, we show that the policy value can be identified with the existence of the learnable value bridge functions only. However, similar to Theorem 2, we would need to introduce the completeness assumption [1]. We summarize the results in the following theorem

Theorem 4 (Identification with completeness)**.** *Suppose a learnable value bridge function b_V exists, and the completeness assumption holds:*

$$\mathbb{E}[g(S, A) \mid A, O^-] = 0 \implies g(S, A) = 0. \tag{7}$$

Then, we have

$$J(\pi^e) = \mathbb{E}_{\tilde{o}\sim\nu_{\mathcal{O}}}[\sum_a b_V(a, \tilde{o})].$$

Compared to Theorem 3, the existence of weight bridge function is replaced with the completeness assumption in Theorem 4.

4.2 Estimation

We first propose three value estimators given certain consistently estimated learnable bridge functions $\hat{b}_V$ and $\hat{b}_W$. We next introduce minimax estimation methods for b_V and b_W.

Given some $\hat{b}_V$ and $\hat{b}_W$, we can construct the following IS and value function-based estimators accordingly to Theorem 3. Specifically, define $\hat{J}_{\mathrm{IS}}$ and $\hat{J}_{\mathrm{VM}}$ to be

$$\frac{1}{1-\gamma}\mathbb{E}_{\mathcal{D}}[\hat{b}_W(A, O^-)R\pi(A \mid O)] \text{ and } \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_a \hat{b}_V(a, \tilde{O})],$$

respectively, where $\mathbb{E}_{\mathcal{D}}$ denotes the empirical average over all observed data tuples, e.g., $\{(O_{i,t-1}, O_{i,t}, A_{i,t}, R_{i,t}, O_{i,t+1})\}_{i,t}$ where i indexes the ith trajectory and t indexes the tth time point.

[1] It is different from Bellman completeness.

In addition, we can combine the aforementioned two estimators for policy evaluation. This yields the following doubly-robust estimator, $\hat{J}_{\mathrm{DR}} = \mathbb{E}_{\mathcal{D}}[J(\hat{b}_W, \hat{b}_V)]$ where

$$J(f,g) \coloneqq \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_a g(a,\tilde{O})] + \frac{f(A,O^-)}{1-\gamma} \times \Big[\{R + \gamma \sum_{a'} g(a', O^+)\}\pi^e(A \mid O) - g(A,O)\Big], \tag{8}$$

for $f : \mathcal{A} \times \mathcal{O} \to \mathbb{R}$ and $g : \mathcal{A} \times \mathcal{O} \to \mathbb{R}$. It has the desired doubly-robust property, as shown in Section 4.3.2.

In fully observable MDPs, $\hat{J}_{\mathrm{IS}}$ reduces to the marginal IS estimator Liu et al. (2018), and $\hat{J}_{\mathrm{DR}}$ reduces to the doubly-robust estimator (Kallus and Uehara, 2019). We remark that our proposal is not a trivial extension of these existing methods to the POMDP setting. The major challenge lies in developing identification formulas in Theorem 3 to correctly identify the target policy's value. These results are not needed in settings without unmeasured confounders.

Next, we present our proposal to estimate b_V and b_W. To estimate b_V, a key observation is that, it satisfies the following integral equation, $\mathbb{E}[L_{\mathbb{V}}(b_V, f)] = 0$ for any f where $L_{\mathbb{V}}(g, f)$ is given by

$$\{R\pi^e(A|O) + \gamma \sum_{a'} g(a', O^+)\pi^e(A|O) - g(A,O)\} f(A, O^-).$$

This motivates us to develop the following minimax learning methods. Specifically, we begin with two function classes $\mathcal{V}, \mathcal{V}^\dagger \subset \{\mathcal{A} \times \mathcal{O} \to \mathbb{R}\}$ and a regularization parameter $\lambda \in \mathbb{R}^+$. We next define the following minimax estimator,

$$\hat{b}_V = \arg\min_{g\in\mathcal{V}} \max_{f\in\mathcal{V}^\dagger} \mathbb{E}_{\mathcal{D}}[L_{\mathbb{V}}(g,f)] - \lambda \mathbb{E}_{\mathcal{D}}[f^2]. \tag{9}$$

Here, we use the function class $\mathcal{V}$ to model the oracle value bridge function, and the function class $\mathcal{V}^\dagger$ to measure the discrepancy between a given $g \in \mathcal{V}$ and the oracle learnable value bridge function. In practice, we can use linear basis functions, neural networks, random forests, reproducing kernel Hilbert spaces (RKHSs), etc., to parameterize these functions. In the fully observable MDP setting, the above optimization reduces to Modified Bellman Residual Minimization (MBRM) when $\lambda = 0.5$ Antos et al. (2008) and Minimax Q-learning (MQL) when $\lambda = 0$ Uehara et al. (2020). It is worthwhile to mention that fitted Q-iteration (FQI, Ernst et al., 2005), a popular policy learning method in MDPs, cannot be straightforwardly extended to the POMDP setting. This is because the regression estimator in each iteration of FQI will be biased under the POMDP setting.

Similarly, according to (6), we consider the following estimator for the weight bridge function,

$$\hat{b}_W = \arg\min_{g\in\mathcal{W}} \max_{f\in\mathcal{W}^\dagger} \mathbb{E}_{\mathcal{D}}[L_{\mathbb{W}}(g,f)] - \lambda \mathbb{E}_{\mathcal{D}}[f^2], \tag{10}$$

for some function classes $\mathcal{W}, \mathcal{W}^\dagger \subset \{\mathcal{A} \times \mathcal{O} \to \mathbb{R}\}$ and some $\lambda \in \mathbb{R}^+$.To simplify the calculation, we can use linear basis functions or RKHSs to parametrize these functions. We discuss this further in Appendix C.

4.3 Theoretical Results for Minimax Estimators

We first investigate the nonasymptotic properties of the value function-based estimator $\hat{J}_{\mathrm{VM}}$. We next study the asymptotic property of the DR estimator $\hat{J}_{\mathrm{DR}}$. Results of the importance sampling estimator can be similarly derived. We discuss this in Appendix C. Our results extend some recently established OPE theories in MDPs (Munos and Szepesvári, 2008; Chen and Jiang, 2019; Kallus and Uehara, 2019; Uehara et al., 2021) to the POMDP setting. We assume the observed dataset $\mathcal{D}$ consist of n i.i.d. copies of (O^-, O, A, R, O^+). This i.i.d. assumption is commonly employed in the RL literature to simplify the proof (see e.g., Dai et al., 2020).

4.3.1 Value function-based methods

We begin with the value function-based estimator. First, to measure the discrepancy between the estimated learnable value bridge function $\hat{b}_V$ and b_V, we introduce the Bellman residual error for POMDPs as follows:

Definition 3. *The Bellman residual operator $\mathcal{T}$ maps a given function $f : \mathcal{A} \times \mathcal{O} \to \mathbb{R}$ to another function $\mathcal{T}f$ such that*

$$\mathcal{T}f(a,o) = \mathbb{E}[R\pi^e(A \mid O) + \gamma \sum_{a'} f(a', O^+)\pi^e(A \mid O) \\ - f(A,O) \mid A = a, O^- = o], \quad \forall a, o.$$

The bellman residual error is defined as $\mathbb{E}[\{\mathcal{T}f(A,O^-)\}^2]$.

By definition, the Bellman residual error is zero for any learnable bridge function. As such, it quantifies how a given function f deviates from the oracle value bridge functions. In MDPs, it reduces to the standard Bellman residual error.

We next establish the rate of convergence of $\hat{b}_V$ in the following theorem.

Theorem 5 (Convergence rate of $\hat{b}_V$)**.** *Set $\lambda > 0$ in Equation 9. Suppose (a) there exists certain learnable bridge function that satisfies Equation 5; (b) $\mathcal{V}$ and $\mathcal{V}^\dagger$ are finite hypothesis classes; (c) $\mathcal{V}$ contains at least one learnable bridge function; (d) $\mathcal{T}\mathcal{V} \subset \mathcal{V}^\dagger$; (e) There exist some constants $C_\mathcal{V}$ and $C_{\mathcal{V}^\dagger}$ such that $\forall v \in \mathcal{V}, \|v\|_\infty \leq C_\mathcal{V}$ and $\forall v \in \mathcal{V}^\dagger, \|v\|_\infty \leq C_{\mathcal{V}^\dagger}$ Then, there exists some universal constant $c > 0$ such that for any $\delta > 0$, with probability $1-\delta$, $\mathbb{E}[\{\mathcal{T}\hat{b}_V(A,O^-)\}^2]^{1/2}$ is upper bounded by*

$$c \max(1, C_\mathcal{V}, C_{\mathcal{V}^\dagger}) \sqrt{\log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)/n}.$$

It is important to note that we only assume the existence of the learnable value bridge function in (a). We do not impose any assumptions on weight bridge functions. Assumption (b) can be further relaxed by assuming that $\mathcal{V}$ and $\mathcal{V}^\dagger$ are general hypothesis classes. In that case, the convergence rate will be characterized by the critical radii of function classes constructed by $\mathcal{V}$ and $\mathcal{V}^\dagger$. See e.g., Uehara et al. (2021) for details. Assumption (c) is the realizability assumption. Since learnable bridge functions are not unique, we only require $\mathcal{V}$ to contain one of them. Assumption (d) is the (Bellman) closedness assumption. It requires that the discriminator class $\mathcal{V}^\dagger$ is sufficiently rich and the operator $\mathcal{B}$ is sufficiently smooth. These assumptions are valid in several examples, including tabular and linear models. To save space, we relegate the related discussions to Appendix C.

Next, we derive the convergence guarantees of the policy value estimator.

Theorem 6 (Convergence rate of $\hat{J}_{\mathrm{VM}}$)**.** *Suppose the weight bridge functions $b'_W(\cdot)$ and learnable value bridge functions exist, and Assumption (a)-(e) in Theorem 5 hold. Then, there exists some universal constant $c > 0$ such that for any $\delta > 0$, with probability $1-\delta$, $|J(\pi^e) - \hat{J}_{\mathrm{VM}}|$ is upper bounded by*

$$\tfrac{c}{1-\gamma}\{\max(1, C_\mathcal{V}, C_{\mathcal{V}^\dagger})\mathbb{E}[b'^2_W(A,O^-)]\}^{\frac{1}{2}} \sqrt{\tfrac{\log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)}{n}}.$$

Theorem 6 uses the identification result in Theorem 3. It requires a stronger condition than Theorem 5, as we assume the existence of weight bridge function. Its proof relies on the following key equation,

$$|J(\pi^e) - \hat{J}_{\mathrm{VM}}| \leq \mathbb{E}[b'^2_W(A,O^-)]^{1/2}\mathbb{E}[\{\mathcal{T}\hat{b}_V(A,O^-)\}^2]^{1/2},$$

where the upper bound for the second term on the right-hand-side is given in Theorem 5.

Next, we present a similar result built upon the identification result in Theorem 4 instead of Theorem 3.

Theorem 7 (Convergence rate of $\hat{J}_{\mathrm{VM}}$ with completenesss)**.** *Suppose learnable value bridge functions exist, the completeness assumption (7) and Assumption (a)-(e) in Theorem 5 hold. Then, there exists some universal constant $c > 0$ such that for any $\delta > 0$, with probability $1 - \delta$, $|J(\pi^e) - \hat{J}_{\mathrm{VM}}|$ is upper bounded by*

$$\frac{c}{1-\gamma}\left(\max(1, C_{\mathcal{V}}, C_{\mathcal{V}^\dagger})\mathbb{E}\left[\left\{\frac{w(S)}{P_{\pi_b}(A|S)}\right\}^2\right]\kappa\right)^{\frac{1}{2}}\sqrt{\frac{\log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)}{n}}, \kappa^2 := \sup_{f\in\mathcal{F}}\frac{\mathbb{E}[\{\mathcal{T}_S f(A,S)\}^2]}{\mathbb{E}[\{\mathcal{T} f(A,O^-\}^2]}$$

where

$$\mathcal{T}_S : [\mathcal{A}\times\mathcal{O}\to\mathbb{R}] \ni f(\cdot) \mapsto \mathbb{E}[R\pi^e(A \mid O) + \sum_{a'} f(a', O)\pi^e(A \mid O) - f(A, O) \mid A = \cdot, S = \cdot] \in [\mathcal{A}\times\mathcal{S}\to\mathbb{R}].$$

Theorem 7 uses the identification result in Theorem 4. Its proof relies on the following key equation:

$$|J(\pi^e) - \hat{J}_{\mathrm{VM}}| \le \mathbb{E}\left[\left\{\frac{w(S)}{P_{\pi_b}(A \mid S)}\right\}^2\right]^{1/2}\mathbb{E}[\{\mathcal{T}_S\hat{b}_V(A, O^-)\}^2]^{1/2}.$$

To translate errors $\mathbb{E}[\{\mathcal{T}_S\hat{b}_V(A, O^-)\}^2]$ into $\mathbb{E}[\{\mathcal{T}\hat{b}_V(A, S)\}^2]$, we introduce a coefficient κ. In linear models, κ is reduced to the ratio between two condition numbers of some covariance matrices. Compared to Theorem 6, the advantage of using Theorem 7 is that it does not involve weight bridge functions. As we have mentioned, it remains challenging to define weight bridge functions when the evaluation policy is history-dependent. However, one can similarly establish the theoretical properties of the corresponding value function-based estimator (see Section 6), as in Theorem 7. The potential disadvantage is that we need to rely on the completeness assumption, which is relatively strong in the non-tabular setting.

Notice that Theorem 5 relies on the closedness assumption, e.g., Assumption (d). While this condition is necessary to obtain the nonasymptotic convergence rate of $\hat{b}_V, \hat{b}_W$, it is *not* necessary to derive the convergence rate of the final policy value estimator. In Theorem 16 (see Appendix C), we show that when $b_V \in \mathcal{V}, b_W \in \mathcal{V}^\dagger$, similar results can be established *without* any closedness conditions.

4.3.2 DR methods

We focus on the DR estimator in this section. We first establish its doubly-robustness property in Theorem 8. It implies that as long as either $\hat{b}_W$ or $\hat{b}_V$ is consistent, the final estimator $\hat{J}_{\mathrm{DR}}$ is consistent.

Theorem 8 (Doubly-robustness property)**.** *Suppose Assumption 2 holds. Then $J(f, g)$ defined in Equation 8 equals $J(\pi^e)$ as long as either $f = b_W$ or $g = b_V$.*

We next show that $\hat{J}_{\mathrm{DR}}$ is efficient in the sense that it is asymptotically normal with asymptotic variance equal to the Cramér-Rao Lower Bound. To simplify the proof, we focus on the tabular setting where $\mathcal{S}, \mathcal{O}$ are discrete spaces.

Theorem 9. *Suppose* $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{O} \mid A = a, \mathbf{O}^-)) = |\mathcal{S}| = |\mathcal{O}|$. *The Cramér-Rao Lower Bound is given by*

$$V_{\mathrm{EIF}} = \mathbb{E}[J(b_W, b_V)^2].$$

We impose a rank assumption in Theorem 9. This condition implies that the bridge functions are uniquely defined, and so are the learnable bridge functions. We remark that without such a uniqueness assumption, the Cramér-Rao Lower Bound is not well-defined.

To avoid imposing certain Donsker conditions (van der Vaart, 1998), we focus on a sample-split version of $\hat{J}_{\mathrm{DR}}$ as in Zheng and van der Laan (2011). It splits all data trajectories in $\mathcal{D}$ into two independent subsets $\mathcal{D}_1, \mathcal{D}_2$, computes the estimated bridge function $\hat{b}_V^{(1)}, \hat{b}_W^{(1)}$ ($\hat{b}_V^{(2)}, \hat{b}_W^{(2)}$) based on the data subset in $\mathcal{D}_2$ ($\mathcal{D}_1$), and does the

estimation of the value based on the remaining dataset. Finally, we aggregate the resulting two estimates to get full efficiency. This yields the following DR estimator,

$$\hat{J}^*_{\mathrm{DR}} = 0.5\mathbb{E}_{\mathcal{D}_1}[J(\hat{b}^{(1)}_W, \hat{b}^{(1)}_V)] + 0.5\mathbb{E}_{\mathcal{D}_2}[J(\hat{b}^{(2)}_W, \hat{b}^{(2)}_V)].$$

We summarize the results in the following theorem.

Theorem 10. *Assume the existence and uniqueness of bridge and learnable bridge functions. Suppose* $\hat{b}_W, b_W, \hat{b}_V, b_V$ *are uniformly bounded by some constants,* $\mathbb{E}[\{\hat{b}^{(j)}_W - b_W\}^2(A, O^-)] = o_p(1), \mathbb{E}[\{\mathcal{T}\hat{b}^{(j)}_V(A, O^-)\}^2] = o_p(1)$, $\mathbb{E}[\{\sum_a \hat{b}^{(j)}_V(a, O) - b_V(a, O)\}^2]^{1/2} = o_p(1), \mathbb{E}[\{\hat{b}^{(j)}_W - b_W\}^2(A, O^-)]^{1/2}\mathbb{E}[\{\mathcal{T}\hat{b}^{(j)}_V(A, O^-)\}^2]^{1/2} = o_p(n^{-1/2})$ *for* $j = 1, 2$. *Then,* $\sqrt{n}(\hat{J}^*_{\mathrm{DR}} - J(\pi^e))$ *weakly converges to a normal distribution with mean 0 and variance* V_{EIF}.

We again, make a few remarks. First, although the estimated bridge functions are required to satisfy certain nonparametric rates only, the resulting value estimator achieves a parametric rate of convergence (e.g. $n^{-1/2}$). This is essentially due to the doubly-robustness property, which ensures that the bias of the estimator can be represented as a product of the difference between the two estimated bridge functions and their oracle values. In comparison, Theorem 6 requires the estimated bridge function to satisfy the parametric convergence rate. This is an advantage of the DR estimator over the value function-based estimator. Second, Theorem 10 derives the asymptotic variance of $\hat{J}^*_{\mathrm{DR}}$, which can be consistently estimated from the data. This allows us to perform a handy Wald-type hypothesis testing. Finally, Theorem 10 requires a stronger assumption, i.e., the uniqueness of bridge functions, which implies that $|\mathcal{S}| = |\mathcal{O}|$. Without this assumption, it remains unclear how to define the L_2 error of the estimated bridge function.

5 OPE in Time-inhomogeneous POMDPs

We consider the time-inhomogeneous setting in this section where the system dynamics, evaluation and behavior policies are allowed to vary over time. We first introduce the identification method for the target policy's value by introducing value and weight bridge functions. We next present the proposed value function-based, IS and DR estimators. We remark that although we focus on evaluation of Markovian policies in this section, the proposed value function-based estimator can be extended to settings where evaluation policies are history-dependent in Section 6.

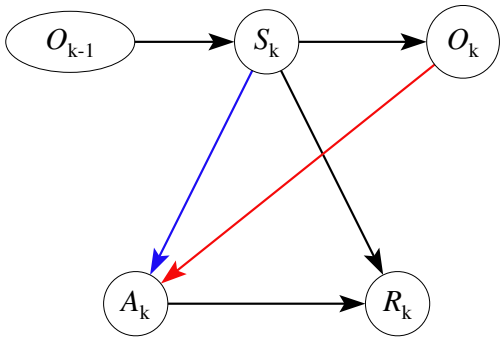
5.1 Identification

We first introduce the bridge function and the learnable bridge function. For any $j \geq 1$, Let $\mu_j(s_j) = \frac{P_{\pi^e,j}(s_j)}{P_{\pi^b,j}(s_j)}$ where $P_{\pi^e,j}$ and $P_{\pi^b,j}$ are the marginal density functions of S_j under π^e and π^b, respectively. Let $\mathfrak{J}_{j-1} = \{O_0, A_0, O_1, A_1, R_1, O_2, \cdots, R_{j-1}\}$.

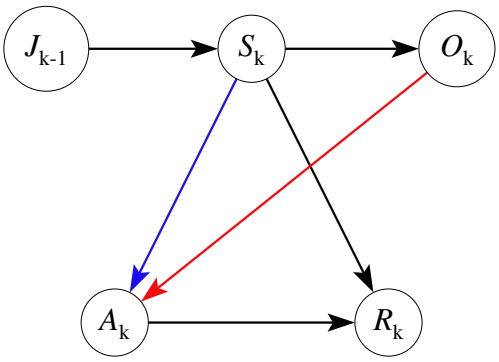
Assumption 3 (Existence of bridge functions)**.** *There exist value bridge functions* $\{b^{\prime[j]}_V\}_{j=0}^{H-1}$ *and weight bridge functions* $\{b^{\prime[j]}_W\}_{j=0}^{H-1}$*, defined as solutions to*

$$\mathbb{E}[b^{\prime[j]}_V(A_j, O_j) \mid A_j, S_j] = \mathbb{E}_{\pi^e}[\sum_{t=j}^{H-1} \gamma^{t-j} R_t \pi^e_j(A_j \mid O_j) | A_j, S_j], \tag{11}$$

$$\mathbb{E}[b^{\prime j}_W(A_j, \mathfrak{J}_{j-1}) \mid A_j, S_j] = \mu_j(S_j) \frac{1}{P_{\pi^b} A_j \mid S_j)}. \tag{12}$$



(a) The case where we use the prior observation as a negative control.



(b) The case where we use the prior history as a negative control.

Definition 4 (Learnable bridge functions)**.** *Learnable value bridge functions* $\{b_V^{[j]}\}_{j=0}^{H-1}$ *and learnable weight bridge functions* $\{b_W^{[j]}\}_{j=0}^{H-1}$ *are defined as solutions to*

$$\begin{aligned}
&\mathbb{E}[b_V^{[j]}(A_j, O_j) \mid A_j, \mathfrak{J}_{j-1}] = \mathbb{E}[R_j \pi_j^e(A_j \mid O_j) + \gamma \sum_{a'} b_V^{[j+1]}(a', O_{j+1}) \pi_j^e(A_j \mid O_j) \mid A_j, \mathfrak{J}_{j-1}], \\
&\mathbb{E}\left[\sum_{a'} b_W^{[j]}(A_j, \mathfrak{J}_{j-1}) \pi_j^e(A_j | O_j) f(O_{j+1}, a') - b_W^{[j+1]}(A_{j+1}, \mathfrak{J}_j) f(O_{j+1}, A_{j+1})\right] = 0, \ \forall f : \mathcal{A} \times \mathcal{O} \to \mathbb{R}.
\end{aligned} \tag{13}$$

It is important to note that we use the whole history $\mathfrak{J}_{j-1}$ in the integral equations to define the learnable bridge functions. Alternatively, one can replace $\mathfrak{J}_{j-1}$ with the most recent observation O_{j-1}. The advantage of using the whole history over a single observation is that it requires weaker assumptions to ensure the existence of the bridge functions. See the discussion below Assumption 4 for details.

Next, we present our key identification theorem under Assumption 3.

Theorem 11. *Suppose Assumption 3 holds. Then, bridge functions are learnable bridge functions. In addition, any learnable bridge function satisfies*

$$J(\pi^e) = \mathbb{E}[\sum_a b_V^{[0]}(a, O_0)], \quad J(\pi^e) = \mathbb{E}[\sum_{t=0}^{H-1} b_W^{[t]}(A_t, \mathfrak{J}_{t-1}) \pi_t^e(A_t \mid O_t) \gamma^t R_t].$$

The next theorem states we can similarly identify the policy value by assuming the completeness.

Theorem 12. *Suppose the existence of leanrbale value bridge functions and the completeness:*

$$\mathbb{E}[g(S_j, A_j) \mid A_j, \mathfrak{J}_{j-1}] = 0 \implies g() = 0.$$

Then, any learnable bridge function satisfies

$$J(\pi^e) = \mathbb{E}[\sum_a b_V^{[0]}(a, O_0)].$$

To further elaborate Assumptions in Theorem 11 and Theorem 12, we focus on the tabular setting in the rest of this section when $\mathcal{S}, \mathcal{O}$, and the reward space are discrete.

First, we see sufficient conditions to ensure Assumptions in Theorem 11. Assumptions in Theorem 11, i.e., Assumption 3, are immediately implied by the following conditions.

Assumption 4. *For any* $0 \leq j \leq H-1$,

$$\mathrm{rank}(\Pr(\mathbf{O}_j \mid \mathbf{S}_j)) = |\mathcal{S}|, \quad \mathrm{rank}(\Pr(\boldsymbol{\mathcal{J}}_{j-1} \mid \mathbf{S}_j)) = |\mathcal{S}|.$$

Next, we see sufficient conditions to ensure Assumption in Theorem 12. Assumptions in Theorem 12 are immediately implied by the following conditions.

Assumption 5. *For any* $0 \leq j \leq H-1$,

$$\mathrm{rank}(\Pr(\mathbf{O}_j \mid \mathbf{S}_j)) = |\mathcal{S}|, \quad \mathrm{rank}(\Pr(\mathbf{S}_j \mid \boldsymbol{\mathcal{J}}_{j-1}, A_j = a)) = |\mathcal{S}|.$$

These assumptions imply that the state space is sufficiently informative and $|\mathcal{O}| \geq |\mathcal{S}|$. Clearly, this shows the advantage of using the whole $\boldsymbol{\mathcal{J}}_{j-1}$ as a negative control rather than O_{j-1} itself since $\mathrm{rank}(\Pr(\boldsymbol{\mathcal{J}}_{j-1} \mid \mathbf{S}_j)) = |\mathcal{S}|$ is weaker than $\mathrm{rank}(\Pr(\mathbf{O}_{j-1} \mid \mathbf{S}_j)) = |\mathcal{S}|$, and similarly, $\mathrm{rank}(\Pr(\mathbf{S}_j \mid \boldsymbol{\mathcal{J}}_{j-1}, A_j = a)) = |\mathcal{S}|$ is weaker than $\mathrm{rank}(\Pr(\mathbf{S}_j \mid \mathbf{O}_{j-1})) = |\mathcal{S}|$.

In addition, Assumption 5 is equivalent to Assumption 2 in Nair and Jiang (2021) ($\mathrm{rank}(\Pr(\mathbf{O}_j \mid \boldsymbol{\mathcal{J}}_{j-1}, A_j = a)) = |\mathcal{S}|$) as we see in the bandit setting. Under this assumption, the policy value is identifiable, as we show in the following lemma.

Lemma 4. *Suppose Assumption 5 holds. Then the policy value equals*

$$\sum_{s_0, a_0, \cdots, r_{H-1}} \sum_{k=0}^{H-1} \gamma^k r_k \tau_k \Pr_{\pi^b}(r_k, o_k \mid a_k, \boldsymbol{\mathcal{J}}_{t-1}) \Pr_{\pi^b}(\mathbf{O}_k \mid a_k, \boldsymbol{\mathcal{J}}_{t-1})^+ \Pr_{\pi^b}(\mathbf{O}_k, o_{k-1} \mid a_{k-1}, \boldsymbol{\mathcal{J}}_{k-2}) \Pr_{\pi^b}(\mathbf{O}_{k-1} \mid a_{k-1}, \boldsymbol{\mathcal{J}}_{k-2})^+ \times \cdots \Pr_{\pi^b}(\mathbf{O}_0),$$

where $\tau_k = \prod_{t=1}^k \pi_t^e(a_t \mid o_t)$.

5.2 Estimation

We present the estimation method in this section. Suppose we have certain consistent estimators $\hat{b}_V = \{\hat{b}_V^{[j]}\}$ and $\hat{b}_W = \{\hat{b}_W^{[j]}\}$ for $b_V = \{b_V^{[j]}\}$ and $b_W = \{b_W^{[j]}\}$, respectively. Theorem 11 suggests that we can estimate the policy value based on the following value function-based and IS estimators.

$$\hat{J}_{\mathrm{VM}} = \mathbb{E}[\sum_a \hat{b}_V^{[0]}(a, O_0)], \quad \hat{J}_{\mathrm{IS}} = \mathbb{E}_{\mathcal{D}}[\sum_{t=0}^{H-1} \hat{b}_W^{[t]}(A_t, \mathfrak{J}_{t-1}) \pi_t^e(A_t \mid O_t) \gamma^t R_t].$$

In addition, we can similarly combine these two estimators to construct the DR estimator $\hat{J}_{\mathrm{DR}} = \mathbb{E}_{\mathcal{D}}[J(\hat{b}_V, \hat{b}_W)]$ where

$$J(f, g) = \sum_a f^{[0]}(a, O_0) + \sum_{k=0}^{H-1} \gamma^k g^{[k]}(A_k, \mathfrak{J}_{k-1}) \left(\pi_k^e(A_k \mid O_k) \{R_k + \gamma \sum_{a^+} f^{[k+1]}(a^+, O_{k+1})\} - f^{[k]}(A_k, O_k) \right),$$

where $f = \{f^{[t]}\}$ and $g = \{g^{[t]}\}$ such that $f^{[t]} : \mathcal{A} \times \mathcal{O} \to \mathbb{R}, g^{[t]} : \mathcal{A} \times \mathcal{J}_{t-1} \to \mathbb{R}$. Here, $\mathcal{J}_t$ denotes the domain over $\mathfrak{J}_t$.

In the fully observable MDP setting, $\hat{J}_{\mathrm{IS}}$ reduces to the one in Xie et al. (2019) and $\hat{J}_{\mathrm{DR}}$ reduces to the one in Kallus and Uehara (2020). The following theorem proves the doubly-robustness property of $\hat{J}_{\mathrm{DR}}$.

Theorem 13 (Doubly robust property). *Under Assumption 3, for any* $f = \{f^{[t]}\}$ *and* $g = \{g^{[t]}\}$ *such that* $g^{[t]} : \mathcal{A} \times \mathcal{J}_{t-1} \to \mathbb{R}$, $f^{[t]} : \mathcal{A} \times \mathcal{O} \to \mathbb{R}$, *we have*

$$J(\pi^e) = \mathbb{E}[J(b_V, g)] = \mathbb{E}[J(f, b_W)].$$

Finally, we discuss how to estimate the learnable bridge functions. Similar to equation 9 and equation 10, we can employ minimax learning to estimate b_V and b_W based on the integral equations in equation 13. However, different from equation 9 and equation 10, b_V and b_W need to be estimated in a recursive fashion in the time-inhomogeneous setting. Specifically, to estimate b_V, we begin by defining $\hat{b}_V^{[H]} = 0$. We next sequentially estimate $\hat{b}_V^{[H-1]}, \hat{b}_V^{[H-2]}, \cdots, \hat{b}_V^{[0]}$ in a backward manner. Similarly, to learn b_W, we define $\hat{b}_W^{[-1]} = 1$ and recursively estimate $\hat{b}_W^{[0]}, \hat{b}_W^{[1]}, \cdots, \hat{b}_W^{[H-1]}$ in a forward manner. Theoretical properties of these minimax estimators can be similarly established, as in Section 4.3, and we omit the technical details.

6 Evaluation of History-Dependent Policies

We have so far assumed that the evaluation policies are Markovian. In this section, we consider the case when evaluation policies are history-dependent, i.e., $\pi^e : \mathcal{O} \times \tilde{\mathcal{J}}_{t-1} \to \Delta(\mathcal{A})$ where $\tilde{\mathcal{J}}_{t-1} = \prod_{k=0}^{t-1}(\mathcal{O} \times \mathcal{A})$. Note $\tilde{\mathcal{J}}_{t-1}$ is different from $\mathcal{J}_{t-1}$ recalling $\mathcal{J}_{t-1} = \mathcal{O} \times \{\prod_{k=0}^{t-1}(\mathcal{O} \times \mathcal{A})\}$. We mainly analyze the value-function based estimator in this section.

We introduce the value and learnable bridge functions as follows. They are natural extensions of Assumption 3 and Definition 4 for history-dependent policies. Let $\tilde{\mathfrak{J}}_{j-1} = \{O_0, A_0, \cdots, O_{H-1}, A_{H-1}\}$. Note this is contained in $\mathfrak{J}_{j-1} = \{O_0^-, O_0, A_0, \cdots, O_{H-1}, A_{H-1}\}$.

Assumption 6 (Existence of bridge functions). *There exist value bridge functions* $\{b_V'^{[j]}\}_{j=0}^{H-1}$ *defined as solutions to*

$$\mathbb{E}[b_V'^{[j]}(A_j, O_j, \tilde{\mathfrak{J}}_{j-1}) \mid A_j, S_j, \tilde{\mathfrak{J}}_{j-1}] = \mathbb{E}_{\pi^e}[\sum_{t=j}^{H-1} \gamma^{t-j} R_t \pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) | A_j, S_j, \tilde{\mathfrak{J}}_{j-1}].$$

There exist weight bridge functions $\{b_W'^{[j]}\}_{j=0}^{H-1}$ *defined as solutions to*

$$\mathbb{E}[b_W'^{[j]}(A_j, \mathfrak{J}_{j-1}) \mid A_j, S_j, \tilde{\mathfrak{J}}_{j-1}] = \mu_j(S_j, \tilde{\mathfrak{J}}_{j-1}) / \pi_b(A_j \mid S_j)$$

where $\mu_j(S_j, \tilde{\mathfrak{J}}_{j-1})$ *is* $P_{\pi^e,j}(S_j, \tilde{\mathfrak{J}}_{j-1}) / P_{\pi^b,j}(S_j, \tilde{\mathfrak{J}}_{j-1})$.

Definition 5 (Learnable bridge functions). *Learnable value bridge functions* $\{b_V^{[j]}\}_{j=0}^{H-1}$ *are defined as solutions to*

$$\mathbb{E}[b_V^{[j]}(A_j, O_j, \tilde{\mathfrak{J}}_{j-1}) \mid A_j, \mathfrak{J}_{j-1}] = \mathbb{E}[R_j \pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) + \gamma \sum_{a'} b_V^{[j+1]}(a', O_{j+1}, \tilde{\mathfrak{J}}_j) \pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) \mid A_j, \mathfrak{J}_{j-1}].$$

It is worthwhile to note $\mathfrak{J}_j$ is contained in $\tilde{\mathfrak{J}}_j$. If evaluation policies depend on the whole $\mathfrak{J}_j$, it is unclear how to identify the policy value in our framework.

Now, we are ready to prove the identification formula.

Theorem 14.

- *Suppose Assumption 6. Then, any learnable bridge function satisfies*

$$J(\pi^e) = \mathbb{E}[\sum_a b_V^{[0]}(a, O_0)].$$

- *Suppose the existence of value bridge functions and the completness assumption:*

$$\mathbb{E}[g(S_j, A_j, \tilde{\mathfrak{J}}_{j-1}) \mid A_j, \mathfrak{J}_{j-1}] = 0 \implies g(\cdot) = 0$$

for any $0 \leq j \leq H-1$. *Then, any learnable bridge function satisfies*

$$J(\pi^e) = \mathbb{E}[\sum_a b_V^{[0]}(a, O_0)].$$

Similar to Section 5, b_V can be estimated in a recursive fashion. In the tabular setting, the sufficient conditions to ensure the existence of value bridge functions and the completness is

$$\text{rank}(\Pr(\mathbf{O}_j \mid A_j = a, \boldsymbol{\mathcal{J}}_{j-1})) = |\mathcal{S}|.$$

The nonasymptotic properties of the resulting estimator can be similarly established as in Section 4.3

7 Experiments

In this section, we evaluate the empirical performance of our method using two synthetic datasets. In both datasets, the state spaces are continuous. Hence, existing POMDP evaluation methods such as Tennenholtz et al. (2020) and Nair and Jiang (2021) are not directly applicable. The discounted factor γ is fixed to 0.95 in all experiments.

7.1 One-Dimensional Dynamic Process

Environment We first consider a simple dynamic process with a one-dimensional continuous state space and binary actions. Similar environments have been considered in the literature (see e.g., Shi et al., 2022). The initial state distribution, reward function and state transition are given by

$$S_0 \sim \mathcal{N}(0, 0.5^2), \quad R_t = S_t + 2A_{t-1} - 1,$$
$$S_{t+1} = 0.5S_t + (2A_t - 1) + \mathcal{N}(0, 0.5^2),$$

respectively. The observation is generated according to the additive noise model, $O_t = S_t + N(0, \sigma_O^2)$. Here, the variance parameter σ_O characterizes the degree of partial observability and hence the degree of unmeasured confounding. In the extreme case where $\sigma_O = 0$, the states become fully-observable and no unmeasured confounders exists. We set the behavior and target policies to be sigmoid functions of the state and the observation, respectively,

$$\pi^b(1|s) = \frac{1}{1+\exp(s+1)}, \pi_w^e(1|o) = \frac{1}{1+\exp(wo+1)},$$

for $w \in \{-3, -2, 1, 2\}$.

Implementation and Baseline Method We use linear basis functions to approximate the bridge functions. In this case, the proposed three estimators ($\hat{J}_{\text{VM}}, \hat{J}_{\text{IS}}, \hat{J}_{\text{DR}}$) coincide with each other (see a similar phenomenon in Uehara et al., 2020), so we compute the value function-based estimator only. We compare our estimator to the standard linear value function-based estimator that assumes no unmeasured confounders, i.e., LSTDQ (Lagoudakis and Parr, 2003). See additional implementation details in Appendix D.

Results For each simulation, we first generate 2000 trajectories with 100 time points per trajectory. This yields a total of 2e5 observations. Reported in the upper panels of Figure 4 are the logarithms of relative mean squared errors of the proposed and baseline methods, with different choices of ω and σ_O. We make a few observations. First, when $\sigma_O \geq 1$, the proposed estimator is significantly better than the baseline estimator. Second, when $\sigma_O = 0.5$, the two methods perform comparably. As we have commented, σ_O measures the degree of unmeasured confounding. For

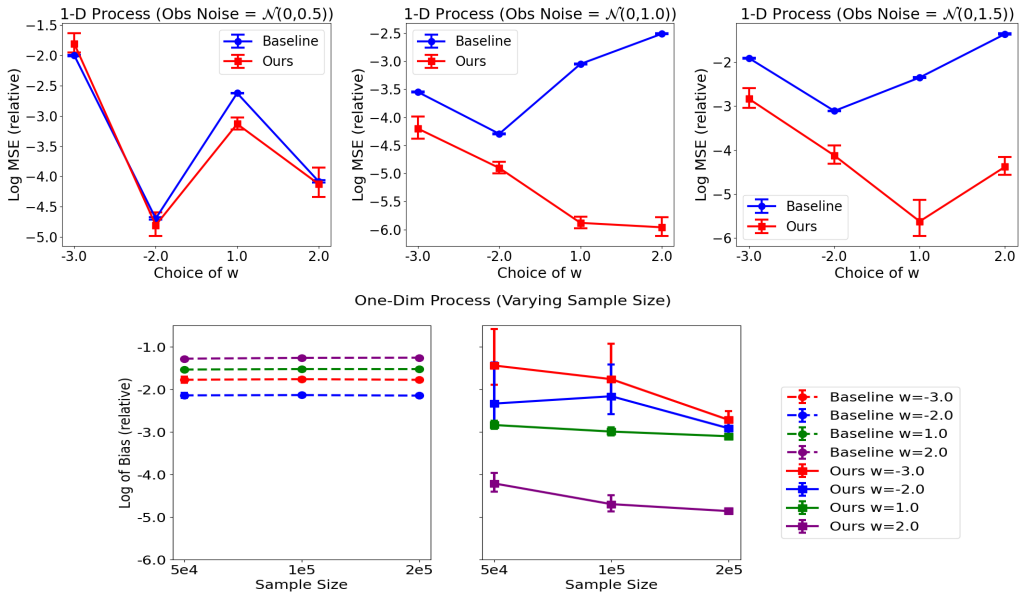


Figure 4: Logarithms of relative biases and mean squared errors (see Appendix D for the detailed definitions) of the proposed and the baseline methods, with difference choices of w and sample sizes. Confidence intervals are calculated from 10 simulations. Upper panels: sample size equals 2e5. σ_O equals 0.5, 1.0 and 1.5, from left to right. Bottom panels: σ_O is fixed to 1.

moderately large σ_O, the baseline estimator cannot handle unmeasured confounders, yielding a biased estimator. In contrast, a small σ_O produces a less confounded dataset. The two estimators thus achieve similar performance.

We next fix σ_O to 1.0, vary the number of trajectories generated in each simulation, and report the corresponding relative biases and mean squared errors of the proposed and baseline methods as well as the associated confidence intervals in the bottom panels of Figure 4 and Figure 6 (see Appendix D). It can be seen that the baseline estimator suffers from a large bias. Their bias and MSE are constant as functions of sample size. To the contrary, the bias and MSE of the proposed estimator decrease with the sample size, demonstrating its consistency.

7.2 CartPole

We next consider the CartPole environment from the OpenAI Gym environment (Brockman et al., 2016). The state variables are 4-dimensional and fully-observable. To create partially observable environments, we similarly simulate observations by adding independent Gaussian noises to each dimension of the states, i.e., $O^{(j)} = S^{(j)}(1 + \mathcal{N}(0, 0.1^2))$, $1 \leq j \leq 4$. To generate data, we apply DQN (Mnih et al., 2015) to the data that include latent states instead of observations, and set the behavior policy to a softmax policy based on the estimated Q-function. The evaluation policy is set to another softmax policy based on DQN applied to the observational data (i.e., no latent states) with the temperature parameter given by τ_O. For each simulation, we collect the dataset according to the behavior policy until the sample size reaches 2e5. We consider three baseline estimators, corresponding to the minimax Q-learning (MQL), minimax weight learning (MWL) and DR estimators (Uehara et al., 2020). These estimators cannot handle unmeasured confounders. The proposed value function-based, marginalized IS and DR estimators are denoted by PO-MQL, PO-MWL and PO-DR, respectively. We parametrize the bridge functions using a two-layer neural network, and set the function spaces $\mathcal{V}^\dagger$ and $\mathcal{W}^\dagger$ to RKHSs to facilitate the computation. Some additional details about the environment and implementation are given in Appendix D. Results are reported in Figure 5. It can be seen that the proposed estimator achieves better performance in all cases.

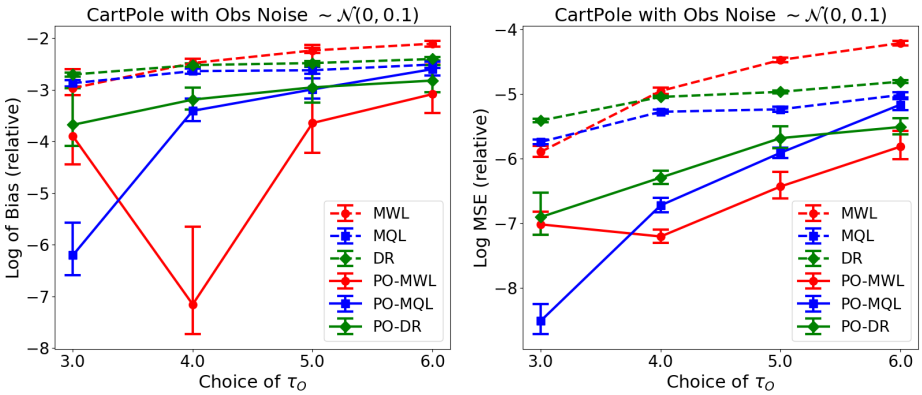


Figure 5: Logarithms of relative biases (left) and MSEs (right) of the proposed (dashed lines) and the baseline (solid lines) estimators and the associated confidence interval, based on 10 simulations, with different choices of the temperature parameter τ_O.

8 Conclusion

We study OPE on confounded POMDPs where behavior policies depend on unobserved state variables. We propose a novel identification method for the target policy's value in the presence of unmeasured confounders. Our proposal only relies on the existence of bridge functions. We further propose minimax learning methods for computing these estimated bridge functions that can be naturally coupled with function approximation to handle a continuous or large state/observation space. We also develop three types of policy value estimators based on the bridge function estimators and provide their nonasymptotic and asymptotic properties. In Section A, we discuss the different between our proposal and a highly-related concurrent work by Bennett and Kallus (2021).

References

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

Antos, A., C. Szepesvári, and R. Munos (2008). Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning 71*(1), 89–129.

Bennett, A. and N. Kallus (2020). The variational method of moments. *arXiv preprint arXiv:2012.09422*.

Bennett, A. and N. Kallus (2021). Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes.

Bennett, A., N. Kallus, L. Li, and A. Mousavi (2021). Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, pp. 1999–2007. PMLR.

Boots, B., S. M. Siddiqi, and G. J. Gordon (2011). Closing the learning-planning loop with predictive state representations. *The International Journal of Robotics Research 30*(7), 954–966.

Brockman, G., V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba (2016). Openai gym. *arXiv preprint arXiv:1606.01540*.

Carrasco, M., J.-P. Florens, and E. Renault (2007). Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics 6*, 5633–5751.

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

Cui, Y., H. Pu, X. Shi, W. Miao, and E. T. Tchetgen (2020). Semiparametric proximal causal inference. *arXiv preprint arXiv:2011.08411*.

Dai, B., O. Nachum, Y. Chow, L. Li, C. Szepesvári, and D. Schuurmans (2020). Coindice: Off-policy confidence interval estimation. *arXiv preprint arXiv:2010.11652*.

Deaner, B. (2018). Proxy controls and panel data. *arXiv preprint arXiv:1810.00283*.

Dikkala, N., G. Lewis, L. Mackey, and V. Syrgkanis (2020). Minimax estimation of conditional moment models. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), *Advances in Neural Information Processing Systems*, Volume 33, pp. 12248–12262.

Ernst, D., P. Geurts, and L. Wehenkel (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research 6*, 503–556.

Feng, Y., L. Li, and Q. Liu (2019). A kernel loss for solving the bellman equation. *arXiv preprint arXiv:1905.10506*.

Futoma, J., M. C. Hughes, and F. Doshi-Velez (2020). Popcorn: Partially observed prediction constrained reinforcement learning. *arXiv preprint arXiv:2001.04032*.

Ghassami, A., A. Ying, I. Shpitser, and E. T. Tchetgen (2021). Minimax kernel machine learning for a class of doubly robust functionals. *arXiv preprint arXiv:2104.02929*.

Hartford, J., G. Lewis, K. Leyton-Brown, and M. Taddy (2017). Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning*, Volume 70, pp. 1414–1423.

Hefny, A., C. Downey, and G. J. Gordon (2015). Supervised learning for dynamical system learning. *Advances in neural information processing systems 28*, 1963–1971.

Hsu, D., S. M. Kakade, and T. Zhang (2012). A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences 78*(5), 1460–1480.

Hu, Y. and S. Wager (2021). Off-policy evaluation in partially observed markov decision processes. *CoRR abs/2110.12343*.

Jiang, N. and L. Li (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pp. 652–661. PMLR.

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

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

Kallus, N., X. Mao, and M. Uehara (2021). Causal inference under unmeasured confounding with negative controls: A minimax learning approach. *arXiv preprint arXiv:2103.14029*.

Kallus, N. and M. Uehara (2019). Efficiently breaking the curse of horizon: Double reinforcement learning in infinite-horizon processes. *arXiv preprint arXiv:1909.05850*.

Kallus, N. and M. Uehara (2020). Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *J. Mach. Learn. Res. 21*, 167–1.

Kallus, N. and A. Zhou (2020). Confounding-robust policy evaluation in infinite-horizon reinforcement learning. In *Advances in Neural Information Processing Systems*, Volume 33, pp. 22293–22304. Curran Associates, Inc.

Kulesza, A., N. Jiang, and S. Singh (2015). Spectral learning of predictive state representations with insufficient statistics. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Kuzborskij, I., C. Vernade, A. Gyorgy, and C. Szepesvári (2021). Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics*, pp. 640–648. PMLR.

Lagoudakis, M. G. and R. Parr (2003). Least-squares policy iteration. *The Journal of Machine Learning Research 4*, 1107–1149.

Levine, S., A. Kumar, G. Tucker, and J. Fu (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*.

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

Liao, P., Z. Qi, and S. Murphy (2020). Batch policy learning in average reward markov decision processes. *arXiv preprint arXiv:2007.11771*.

Liu, Q., L. Li, Z. Tang, and D. Zhou (2018). Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems 31*, pp. 5356–5366.

Mastouri, A., Y. Zhu, L. Gultchin, A. Korba, R. Silva, M. J. Kusner, A. Gretton, and K. Muandet (2021). Proximal causal learning with kernels: Two-stage estimation and moment restriction. *arXiv preprint arXiv:2105.04544*.

Miao, W., Z. Geng, and E. J. T. Tchetgen (2018). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika 105*(4), 987–993.

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

Muandet, K., A. Mehrjou, S. K. Lee, and A. Raj (2019). Dual instrumental variable regression. *arXiv preprint arXiv:1910.12358*.

Munos, R. and C. Szepesvári (2008). Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research 9*(5).

Nachum, O., Y. Chow, B. Dai, and L. Li (2019). Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *arXiv preprint arXiv:1906.04733*.

Nair, Y. and N. Jiang (2021). A spectral approach to off-policy evaluation for pomdps. *arXiv preprint arXiv:2109.10502*.

Namkoong, H., R. Keramati, S. Yadlowsky, and E. Brunskill (2020). Off-policy policy evaluation for sequential decisions under unobserved confounding. *arXiv preprint arXiv:2003.05623*.

Pananjady, A. and M. J. Wainwright (2021). Instance-dependent l_∞-bounds for policy evaluation in tabular reinforcement learning. *IEEE transactions on information theory 67*(1), 566–585.

Precup, D. (2000). Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, 80.

Shi, C., R. Wan, V. Chernozhukov, and R. Song (2021). Deeply-debiased off-policy interval estimation. *arXiv preprint arXiv:2105.04646*.

Shi, C., X. Wang, S. Luo, H. Zhu, J. Ye, and R. Song (2022). Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. *Journal of the American Statistical Association* (just-accepted), 1–29.

Shi, C., J. Zhu, Y. Shen, S. Luo, H. Zhu, and R. Song (2022). Off-policy confidence interval estimation with confounded markov decision process. *arXiv preprint arXiv:2202.10589*.

Singh, R. (2020). Kernel methods for unobserved confounding: Negative controls, proxies, and instruments. *arXiv preprint arXiv:2012.10315*.

Song, L., B. Boots, S. M. Siddiqi, G. Gordon, and A. Smola (2010). Hilbert space embeddings of hidden markov models. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pp. 991–998.

Tennenholtz, G., U. Shalit, and S. Mannor (2020). Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Volume 34, pp. 10276–10283.

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

Uehara, M., J. Huang, and N. Jiang (2020). Minimax weight and q-function learning for off-policy evaluation. In *International Conference on Machine Learning*, pp. 9659–9668. PMLR.

Uehara, M., M. Imaizumi, N. Jiang, N. Kallus, W. Sun, and T. Xie (2021). Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*.

van der Vaart, A. W. (1998). *Asymptotic statistics*. Cambridge, UK: Cambridge University Press.

Wang, L., Z. Yang, and Z. Wang (2020). Provably efficient causal reinforcement learning with confounded observational data. *arXiv preprint arXiv:2006.12311*.

Xie, T., Y. Ma, and Y.-X. Wang (2019). Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. In *Advances in Neural Information Processing Systems 32*, pp. 9665–9675.

Xu, L., H. Kanagawa, and A. Gretton (2021). Deep proxy causal learning and its application to confounded bandit policy evaluation. *arXiv preprint arXiv:2106.03907*.

Yang, M., O. Nachum, B. Dai, L. Li, and D. Schuurmans (2020). Off-policy evaluation via the regularized lagrangian. *arXiv preprint arXiv:2007.03438*.

Yin, M. and Y.-X. Wang (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 3948–3958. PMLR.

Ying, A., W. Miao, X. Shi, and E. J. T. Tchetgen (2021). Proximal causal inference for complex longitudinal studies. *arXiv preprint arXiv:2109.07030*.

Zhang, J. and E. Bareinboim (2016). Markov decision processes with unobserved confounders: A causal approach. Technical report, Technical Report R-23, Purdue AI Lab.

Zhang, J. and E. Bareinboim (2021). Non-parametric methods for partial identification of causal effects. *Columbia CausalAI Laboratory Technical Report (R-72)*.

Zhang, M. R., T. L. Paine, O. Nachum, C. Paduraru, G. Tucker, Z. Wang, and M. Norouzi (2021). Autoregressive dynamics models for offline policy evaluation and optimization. *arXiv preprint arXiv:2104.13877*.

Zheng, W. and M. J. van der Laan (2011). Cross-validated targeted minimum-loss-based estimation. In *Targeted Learning: Causal Inference for Observational and Experimental Data*, Springer Series in Statistics, pp. 459–474. New York, NY: Springer New York.

A Additional Related Works

We discuss some additional related works in this section.

Minimax Estimation Our proposal defines the value and weight bridge functions as solutions to some integral equations. Nonparametrically estimating these bridge functions is closely related to nonparametric instrumental variables (NPIV) estimation. In NPIV estimation, the standard regression estimator is biased and variants of minimax estimation methods have been developed to correct the bias (Dikkala et al., 2020; Muandet et al., 2019; Hartford et al., 2017; Bennett and Kallus, 2020). In the RL literature, the minimax learning method has also been widely used for OPE without unmeasured confounders Antos et al. (2008); Chen and Jiang (2019); Nachum et al. (2019); Feng et al. (2019); Uehara et al. (2020).

Comparison with Bennett and Kallus (2021) The difference between the work of Bennett and Kallus (2021) and our proposal can be summarized as follows. Methodologically, the IS, VM, DR estimators developed in the two papers are *different*. First, consider the IS estimator as an example. In fully-observable MDPs, our proposed estimator is reduced to marginal IS estimators (Liu et al., 2018; Xie et al., 2019) whereas the IS estimator developed by Bennett and Kallus (2021) is reduced to the cumulative IS estimator (Precup, 2000). Second, Bennett and Kallus (2021) do not introduce value bridge functions as in our paper. Instead, they introduce an outcome bridge function to approximate the reward function at each time step $0 \leq t \leq H-1$. Because of the difference between these definitions, their estimating procedure for the outcome bridge functions involves the estimated weight bridge functions. In contrast, the estimation of value bridge functions in our paper is agnostic to the estimation of weight bridge function.

Theoretically, we focus on the derivation of finite-sample error bounds. To the contrary, Bennett and Kallus (2021) focus on establishing asymptotic properties. In addition, the efficiency bound they developed differs from ours (see Theorem 9). Specifically, they focused on the non-tabular case whereas our bounded is limited to the tabular setting. Moreover, since we focus on evaluating Markovian and stationary policies, in fully-observable MDPs, our bounds are reduced to those in Kallus and Uehara (2019) whereas their bound is reduced to the one in Jiang and Li (2016).

B Some Additional Details Regarding OPE in Partially Observable Contextual Bandits

We specialize our results in Section 3 to tabular settings. We use the following notation. Let X, Y be random variables taking values $\{x_1, \cdots, x_n\}$ and $\{y_1, \cdots, y_m\}$, respectively. Then we denote a $n \times m$ matrix with elements $\Pr(x_i \mid y_j)$ by $\Pr(\mathbf{X} \mid \mathbf{Y})$. Similarly, $\Pr(\mathbf{X})$ denotes a n-dimensional vector with elements x_i. We denote the Moore-penrose inverse of $\Pr(\mathbf{X} \mid \mathbf{Y})$ by $\Pr(\mathbf{X} \mid \mathbf{Y})^+$.

We next introduce the following assumption.

Assumption 7. *Suppose* $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{O}_0|\mathbf{S}_0)) = |\mathcal{S}|$ *and* $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{O}_{-1}|\mathbf{S}_0)) = |\mathcal{S}|$.

As we have mentioned, Assumption 7 implies that $|\mathcal{O}| \geq |\mathcal{S}|$ and is weaker than the condition in Tennenholtz et al. (2020) that requires the state and observation spaces have the same cardinality. The following lemma states that 7 is equivalent to Assumption 2 in Nair and Jiang (2021).

Lemma 5. *Suppose overlap conditions:*

$$\Pr_{\pi^b}(O^- = o^- \mid A_0 = a) > 0, \quad \Pr_{\pi^b}(O^- = o^- \mid A_0 = a) > 0$$

for any $(s,a,o) \in \mathcal{S}\times\mathcal{A}\times\mathcal{O}$*. Then, Assumption 7 holds if and only if* $\operatorname{rank}(\Pr_{\pi^b}(\mathbf{O}_0|A_0=a,\mathbf{O}_{-1}))=|\mathcal{S}|$ *for any* $a\in\mathcal{A}$*.*

Then following Nair and Jiang (2021), the policy value can be explicitly identified as follows:

Theorem 15 (Identification formula in the tabular setting)**.**

$$J(\pi^e)=\sum_{a,r,o} r\pi^e(a|o)\Pr_{\pi^b}(r,o|\mathbf{O}_{-1})\{\Pr_{\pi^b}(\mathbf{O}_0|A_0=a,\mathbf{O}_{-1})\}^{+}\Pr_{\pi^b}(\mathbf{O}_0).$$

C Some Additional Details Regarding OPE in Time-Homogeneous POMDPs

As we have commented, we can use linear basis functions or RKHSs to estimate the bridge functions. Specifically, when $\mathcal{V}^\dagger$ or $\mathcal{W}^\dagger$ is set to be the unit ball in an RKHS, then there exists a close form expression for the inner maximization problem (see e.g., Liu et al., 2018). When linear basis functions are used, the proposed minimax optimization has explicit solutions, as we elaborate in the example below.

Example 1 (Linear models)**.** *Let* $\mathcal{V}=\mathcal{V}^\dagger=\{(a,o)\mapsto\theta^\top\phi(a,o):\theta\in\Theta\}$ *and* $\lambda=0$*, where* $\phi:\mathcal{A}\times\mathcal{O}\to\mathbb{R}^d$ *is a feature vector and* $\Theta\subset\mathbb{R}^d$*. When the parameter space* Θ *is sufficiently large, it is immediate to see that*

$$\hat{b}_V=\left(\mathbb{E}_{\mathcal{D}}[\phi(A,O^-)\phi^\top(A,O)]-\gamma\mathbb{E}_{\mathcal{D}}\left[\sum_{a'}\pi^e(A\mid O)\phi(A,O^-)\phi^\top(a',O^+)\right]\right)^{+}\mathbb{E}_{\mathcal{D}}[R\pi^e(A\mid O)\phi(A,O^-)].$$

The resulting policy value estimator is given by

$$\mathbb{E}_{\mathcal{D}}[\sum_{a}\phi(a,O)]^\top\left(\mathbb{E}_{\mathcal{D}}[\phi(A,O^-)\phi^\top(A,O)]-\gamma\mathbb{E}_{\mathcal{D}}\left[\sum_{a'}\pi^e(A\mid O)\phi(A,O^-)\phi^\top(a',O^+)\right]\right)^{+}\mathbb{E}_{\mathcal{D}}[R\pi^e(A\mid O)\phi(A,O^-)].$$

Consider the standard MDP setting where $S=O=O^-, S^+=O^+$*. By setting* $\mathcal{V}=\{(s,a)\mapsto\theta^\top\pi^e(a\mid s)\phi(a,s)\}, \mathcal{V}^\dagger=\{(s,a)\mapsto\theta^\top\phi(a,s)/\pi^e(a\mid s)\}$*, the above estimator reduces to classical LSTDQ estimator Lagoudakis and Parr (2003):*

$$\mathbb{E}_{\mathcal{D}}[\phi(\pi^e,S)]^\top\left(\mathbb{E}_{\mathcal{D}}[\phi(A,S)\phi^\top(A,S)]-\gamma\mathbb{E}_{\mathcal{D}}\left[\phi(A,S)\phi^\top(\pi^e,S^+)\right]\right)^{+}\mathbb{E}_{\mathcal{D}}[R\phi(A,S)],$$

where $\phi(\pi^e,s)=\mathbb{E}_{a\sim\pi^e(s)}[\phi(a,s)]$*.*

Alternatively, we may set $\mathcal{V}:=\{(a,o)\to\pi^e(a|o)\theta^\top\phi(a,o):\theta\in\Theta\}$ *instead. As discussed in Section 7.1, we find such a parametrization has smaller approximation error in the implementation. The corresponding estimator is given by*

$$\mathbb{E}_{O\sim\nu_O,a\sim\pi^e(\cdot|O)}[\phi(a,O)]^\top\big(\mathbb{E}_{\mathcal{D}}[\phi(A,O^-)\phi^\top(A,O)\pi^e(A\mid O) - \gamma\phi(A,O^-)\pi^e(A\mid O)\mathbb{E}_{a'\sim\pi^e(\cdot|O^+)}[\phi^\top(a',O^+)]]\big)^{+}\mathbb{E}_{\mathcal{D}}[R\pi^e(A\mid O)\phi(A,O^-)].$$

We next introduce two examples to further elaborate Theorem 5.

Example 2 (Tabular models)**.** *Consider the tabular case, and* $\mathcal{V}$ *and* $\mathcal{V}^\dagger$ *are fully expressive classes, i.e., linear in the one-hot encoding vector over* $(\mathcal{A}\times\mathcal{O})$*. Then,* $\log(|\mathcal{V}||\mathcal{V}^\dagger|)$ *is substituted by* $|\mathcal{O}||\mathcal{A}|$*. The Bellman closedness* $\mathcal{T}\mathcal{V}\subset\mathcal{V}^\dagger$ *is satisfied. Our finite sample result circumvents the potentially complicated matrix concentration argument in POMDPs (see e.g., Hsu et al., 2012) by viewing the problem from a general perspective.*

Example 3 (Linear models)**.** *Consider the case where $\mathcal{V}$ and $\mathcal{V}^\dagger$ are linear models. Then, $\log(|\mathcal{V}||\mathcal{V}|^\dagger)$ can be substituted by the dimension of the feature vector d. The Bellman closedness assumption is satisfied when $P_{\pi^b}(R, O, O^+ \mid A, O^-)$ is linear in $\phi(A, O^-)$. In fully-observable MDPs, this reduces to the linear MDP model (Jin et al., 2020), i.e., $P_{\pi^b}(R, S^+ \mid A, S)$ is linear in $\phi(A, S)$.*

The next theorem shows that the finite sample rate of convergence for the value function-based estimator can be established without any closedness assumption.

Theorem 16 (Convergence rate without closedness)**.** *Set λ in equations 9 and 10 to zero. Suppose (a) Assumption 2 holds so that the bridge functions exist; (b) $\mathcal{V}$ and $\mathcal{V}^\dagger$ are symmetric finite hypothesis classes; (c) $\mathcal{V}$ contains certain learnable reward bridge function; (d) $\mathcal{V}^\dagger$ contains learnable weight bridge function; (e) $\forall v \in \mathcal{V}, \|v\|_\infty \leq C_{\mathcal{V}}$ and $\forall v \in \mathcal{V}^\dagger, \|v\|_\infty \leq C_{\mathcal{V}^\dagger}$. Then, with probability $1-\delta$,*

$$|J(\pi^e) - \hat{J}_{\mathrm{VM}}| \leq c \max(1, C_{\mathcal{V}}, C_{\mathcal{V}^\dagger})^2 (1-\gamma)^{-2} \sqrt{\log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)/n},$$

for some universal constant $c > 0$.

Next, we discuss the theoretical properties of the IS estimator. To establish its convergence rate, we can define the adjoint Bellman residual operator $\mathcal{T}'$ as in Uehara et al. (2021), and derive the convergence rate of $\mathbb{E}[\{\mathcal{T}'\hat{b}_W(A, O^+)\}^2]^{1/2}$ under realizability ($b_W \in \mathcal{W}$) and (adjoint) closedness ($\mathcal{T}'\mathcal{W} \subset \mathcal{W}^\dagger$). Similar to the proof of Theorem 6, this (adjoint) Bellman residual error can be translated into the error of the estimated policy value. Without the closedness assumption, similar results can be derived when $b_W \in \mathcal{W}$ and $b_V \in \mathcal{W}^\dagger$.

D Experiments

Measure of Estimation Error Given n dataset $D_1, D_2, ..., D_n$, the estimators computed based on each dataset $\hat{V}_1, ..., \hat{V}_n$, and the true value V, we define the relative bias to be

$$|\frac{1}{n}\sum_{i=1}^{n}\frac{\hat{V}_i}{V} - 1|.$$

Define the relative mean squared error to be:

$$|\frac{1}{n}\sum_{i=1}^{n}\Big(\frac{\hat{V}_i - V}{V}\Big)^2|.$$

In our experiments, we use the above two definitions to measure the estimation error of different estimators.

Additional Details for the 1d Continuous Dynamic Process Example Notice that the definition of the bridge value function involves the evaluation policy. Instead of directly using linear models to parametrize b_V, we set the function space to $\mathcal{V} := \{(a, o) \to \pi^e(a|o)\theta^\top\phi(a, o) : \theta \in \Theta\}$. We find that such a parametrization has smaller approximation error. The closed-form expression of the resulting estimator is given in Section C. We use the Python function RBFsampler to generate random Fourier features. To mitigate the randomness arising from the features, for each dataset and each method, we use 5 different random seeds to generate 5 sets of RBF features and use the average value as the final estimator. The RBF kernel is set to $\exp(-5x^2)$, and the feature dimension is set to 100.

Additional Figures We next report the relative MSE of the two estimators in the following figure.

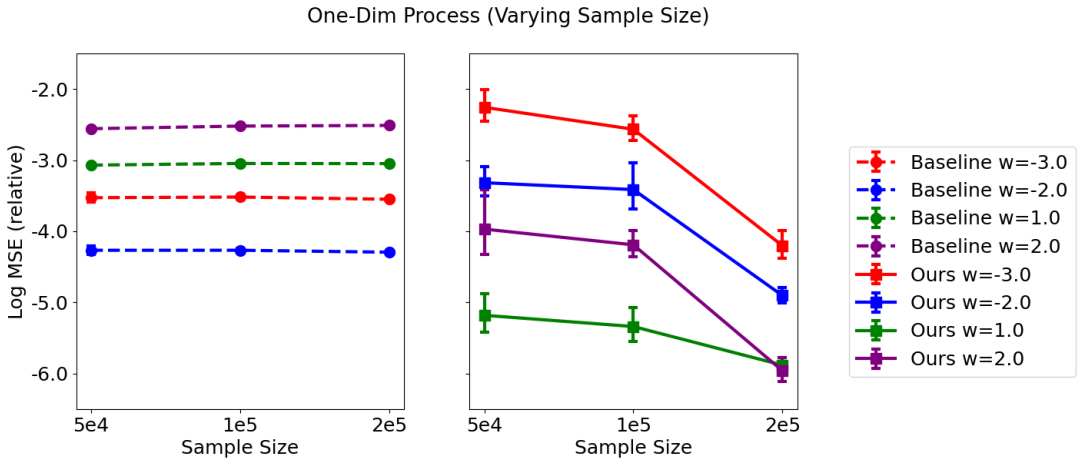


Figure 6: Experiments with varying sample size. Noise Level $\sigma_O = 1.0$

CartPole Environment The state space is 4-dimensional, including position and velocity of cart, and angle and angle velocity of pole. The action space is $\{0, 1\}$, corresponding to pushing the cart to the left or to the right. In addition, we use a modified reward function to better distinguish values among different policies. Specifically, the reward is defined as

$$r = |2.0 - \frac{x}{x_{\text{clip}}}| \cdot |2.0 - \frac{\theta}{\theta_{\text{clip}}}| - 1.0$$

where x and θ are the position of Cart and angle of Pole, respectively, and x_{clip} and θ_{clip} are the thresholds such that the episode will terminate (done = True) when either $|x| \geq x_{clip}$ or $|\theta| \geq \theta_{clip}$. Under this definition, the reward will be larger when the cart is closer to the center and the angle of the pole is closer to perpendicular. It is straightforward to show that r is bounded between 0 and 3. Since we set $\gamma = 0.5$, the value is bounded between 0 and 60.

CartPole Implementation We set the adversarial function spaces $\mathcal{V}^\dagger$ and $\mathcal{W}^\dagger$ to RKHSs to facilitate the computation. Both b_v in PO-MQL and b_w in PO-MWL are parameterized by a two-layer neural network with layer width = 256 and ReLU as activation function, and are optimized by a kernel loss function (see the derivation below).

In the following, we use $K(x_1; x_2)$ to denote the RBF kernel:

$$K(x_1; x_2) := \exp(-\frac{\|x_1 - x_2\|_2}{2\beta^2})$$

where β denotes the bandwidth parameter. We choose $\beta = \text{med}/2$ during the training of PO-MWL (b_w) or MWL, and $\beta = \text{med}/5$ for PO-MQL (b_v) or MQL, where "med" is the median of the l2-distance over the samples in the dataset.

Derivation of the Loss Function for PO-MQL (b_v) Similar to the experiments in Toy environments, we reparameterize g function with $\bar{g} \cdot \pi^e$, and therefore the predictor with g should be adjusted by replacing g with $\bar{g} \cdot \pi^e$:

$$\mathbb{E}_{O \sim \nu_O, A \sim \pi^e}[\bar{g}(O, A)\pi^e(A|O)].$$

The loss function is given by:

$$\begin{aligned}\max_f L_V^2(g,f) =& \max_f \Big(\mathbb{E}[(R\pi^e(A|O)+\gamma\mathbb{E}_{a'\sim\pi^e}[g(a',O^+)]\pi^e(A|O)-g(A,O)\pi^e(A|O))f(A,O^-)]\Big)^2\\ =& \max_f \Big(\Big\langle f, \mathbb{E}[R\pi^e(A|O)+\gamma\mathbb{E}_{a'\sim\pi^e}[g(a',O^+)]\pi^e(A|O)-g(A,O)\pi^e(A|O)]K(A,O^-;\cdot)\Big\rangle_{\mathcal{H}_K}\Big)^2\\ =&\mathbb{E}[\Big(R\pi^e(A|O)+\gamma\mathbb{E}_{a'\sim\pi^e}[g(a',O^+)]\pi^e(A|O)-\pi^e(A|O)g(A,O)\Big)K(A,O^-;\bar{A},\bar{O}^-)\\ &\cdot\Big(\bar{R}\pi^e(\bar{A}|\bar{O})+\gamma\mathbb{E}_{\bar{a}'\sim\pi^e}[g(\bar{a}',\bar{O}^+)]\pi^e(\bar{A}|\bar{O})-\pi^e(\bar{A}|\bar{O})g(\bar{A},\bar{O})\Big)].\end{aligned}$$

Derivation of the Loss Function for PO-MWL (b_w) Similarly, we reparameterize f function with $\bar{f}\cdot\pi^e$. Since we do not change the parameterization of b_w, we only need to adjust the loss function without changing the estimator:

$$\begin{aligned}&\max_f L_W^2(g,f)\\ =&\max_f \Big(\mathbb{E}[\gamma g(O^-,A)\pi^e(A|O)\sum_{a'}f(O^+,a')-g(O^-,A)f(O,A)]+(1-\gamma)\mathbb{E}_{O\sim\nu_O}[\sum_{a'}f(O,a')]\Big)^2\\ =&\max_f \Big(\mathbb{E}[\gamma g(O^-,A)\pi^e(A|O)\mathbb{E}_{a'\sim\pi^e(\cdot|O^+)}[\bar{f}(O^+,a')]-g(O^-,A)\pi^e(A|O)\bar{f}(O,A)]+(1-\gamma)\mathbb{E}_{O\sim\nu_O,a'\sim\pi^e}[\bar{f}(O,a')]\Big)^2\\ =&\max_f \Big(\mathbb{E}[\gamma g(O^-,A)\pi^e(A|O)\mathbb{E}_{a'\sim\pi^e(\cdot|O^+)}[\langle\bar{f},K(O^+,a';\cdot)\mathrm{a}]-g(O^-,A)\pi^e(A|O)\langle\bar{f},K(O,A;\cdot)\mathrm{a}]\\ &+(1-\gamma)\mathbb{E}_{O\sim\nu_O,a'\sim\pi^e}[\langle\bar{f},K(O,a';\cdot)\mathrm{a}]\Big)^2\\ =&\max_f \Big(\langle\bar{f},\mathbb{E}[\gamma g(O^-,A)\pi^e(A|O)\Big(\mathbb{E}_{a'\sim\pi^e(\cdot|O^+)}[K(O^+,a';\cdot)]-K(O,A;\cdot)\Big)]+(1-\gamma)\mathbb{E}_{O\sim\nu_O,a'\sim\pi^e}[K(O,a';\cdot)]\mathrm{a}\Big)^2\\ =&\mathbb{E}[g(O^-,A)\pi^e(A|O)g(\bar{O}^-,\bar{A})\pi^e(\bar{A}|\bar{O})\cdot\\ &\qquad\Big(\gamma^2K(O^+,\pi^e;\bar{O}^+,\pi^e)+K(O,A;\bar{O},\bar{A})-\gamma K(O^+,\pi^e;\bar{O},\bar{A})-\gamma K(O,A;\bar{O}^+,\pi^e)\Big)]\\ &+(1-\gamma)^2\mathbb{E}[K(O_0,\pi^e;\bar{O}_0,\pi^e)] \qquad\qquad\text{(no gradient)}\\ &+\gamma(1-\gamma)\mathbb{E}[g(O^-,A)\pi^e(A|O)K(O^+,\pi^e;\bar{O}_0,\pi^e)+g(\bar{O}^-,\bar{A})\pi^e(\bar{A}|\bar{O})K(O_0,\pi^e;\bar{O}^+,\pi^e)]\\ &-(1-\gamma)\mathbb{E}[g(O^-,A)\pi^e(A|O)K(O,A;\bar{O}_0,\pi^e)+g(\bar{O}^-,\bar{A})\pi^e(\bar{A}|\bar{O})K(O_0,\pi^e;\bar{O},\bar{A})].\end{aligned}$$

where we denote

$$K(O,\pi^e;\bar{O},\pi^e):=\mathbb{E}_{a'\sim\pi^e(\cdot|O),\bar{a}'\sim\pi^e(\cdot|\bar{O})}[K(O,a';\bar{O},\bar{A}),\quad K(O,\pi^e;\bar{O},\bar{A}):=\mathbb{E}_{a'\sim\pi^e(\cdot|O)}[K(O,a';\bar{O},\bar{A})].$$

Loss Function for Baseline Estimators We follow Uehara et al. (2020) to define the MQL and MWL loss function with the adversarial function space given by RHKSs. The loss function for MQL is given by:

$$\max_f L_q^2(f,q) =\mathbb{E}[\Big(R+\gamma\mathbb{E}_{a'\sim\pi^e(\cdot|O^+)}[q(a',O^+)]-q(A,O)\Big)K(A,O;\bar{A},\bar{O})\Big(\bar{R}+\gamma\mathbb{E}_{\bar{a}'\sim\pi^e(\cdot|\bar{O}^+)}[q(\bar{a}',\bar{O}^+)]-q(\bar{A},\bar{O})\Big)].$$

The loss function for MWL is given by:

$$
\begin{aligned}
&\max_f L_w^2(f,w)\\
=&\mathbb{E}[w(O,A)w(\bar{O},\bar{A})\Big(K(O^+,\pi^e;\bar{O}^+,\pi^e)+K(O,A;\bar{O},\bar{A})-\gamma K(O^+,\pi^e;\bar{O},\bar{A})-\gamma K(O,A;\bar{O}^+,\pi^e)\Big)]\\
&+(1-\gamma)^2\mathbb{E}[K(O_0,\pi^e;\bar{O}_0,\pi^e)] && \text{(no gradient)}\\
&+\gamma(1-\gamma)\mathbb{E}[w(O,A)K(O^+,\pi^e;\bar{O}_0,\pi^e)+w(\bar{O},\bar{A})K(O_0,\pi^e;\bar{O}^+,\pi^e)]\\
&-(1-\gamma)\mathbb{E}[w(O,A)K(O,A;\bar{O}_0,\pi^e)+w(\bar{O},\bar{A})K(O_0,\pi^e;\bar{O},\bar{A})].
\end{aligned}
$$

In addition, we use the same neural network architecture and the same choice of bandwidth during the training of MQL/MWL as those for PO-MQL/PO-MWL.

E Omitted proof

E.1 Proof of Section 3

Proof of Lemma 1. First, we prove the value-function based identification formula:

$$
\begin{aligned}
&\mathbb{E}[\sum_a b'_V(a,O_0)]=\mathbb{E}[\sum_a \mathbb{E}[b'_V(a,O_0)|S_0]]=\mathbb{E}[\sum_a \mathbb{E}[b'_V(a,O_0)|A_0=a,S_0]] && (A_0\perp O_0\mid S_0)\\
&=\mathbb{E}[\sum_a \mathbb{E}[R_0\pi^e(a|O_0)|A_0=a,S_0]] && \text{(Definition of bridge functions)}\\
&=\mathbb{E}[\sum_a \mathbb{E}[R_0|A_0=a,S_0]\mathbb{E}[\pi^e(a\mid O_0)\mid A_0=a,S_0]] && (R_0\perp O_0\mid A_0,S_0)\\
&=\mathbb{E}[\sum_a \mathbb{E}[R_0|A_0=a,S_0]\mathbb{E}[\pi^e(a\mid O_0)\mid S_0]] && (A_0\perp O_0\mid S_0)\\
&=\mathbb{E}[\sum_a \mathbb{E}[R_0|A_0=a,S_0]\pi^e(a\mid O_0)]\\
&=J(\pi^e). && \text{(From standard regression formula)}
\end{aligned}
$$

Second, we prove the importance sampling identification formula:

$$
\begin{aligned}
&\mathbb{E}[b_W(A_0,O_0^-)R_0\pi^e(A_0\mid O_0)]=\mathbb{E}[\mathbb{E}[b'_W(A_0,O_0^-)\mid S_0,A_0]R_0\pi^e(A_0\mid O_0)] && (O_0^-\perp R_0\mid S_0,A_0)\\
&=\mathbb{E}[1/\pi^b(A_0\mid S_0)\pi^e(A_0\mid O_0)R_0] && \text{(Definition of } b'_W)\\
&=J(\pi^e). && \text{(From standard IPW formula)}
\end{aligned}
$$

□

Proof of Lemma 2. We first prove reward bridge functions are learnable reward bridge functions:

$$
\begin{aligned}
&\mathbb{E}[R_0\pi^e(A_0|O_0)\mid A_0,O_0^-]=\mathbb{E}[\mathbb{E}[R_0\pi^e(A_0|O_0)\mid A_0,S_0,O_0^-]\mid A_0,O_0^-]\\
&=\mathbb{E}[\mathbb{E}[R_0\pi^e(A_0|O_0)\mid A_0,S_0]\mid A_0,O_0^-] && (R_0,O_0\perp O_0^-\mid S_0,A_0)\\
&=\mathbb{E}[\mathbb{E}[b'_V(A_0,O_0)\mid A_0,S_0]\mid O_0^-,A_0] && \text{(Definition of bridge functions)}\\
&=\mathbb{E}[\mathbb{E}[b'_V(A_0,O_0)\mid S_0,O_0^-,A_0]\mid O_0^-,A_0] && (O_0\perp O_0^-\mid S_0,A_0)\\
&=\mathbb{E}[b'_V(A_0,O_0)\mid O_0^-,A_0].
\end{aligned}
$$

Next, we prove weight bridge functions are learnable weight bridge functions:

$$\begin{aligned}
&\mathbb{E}[b'_W(A_0, O_0^-) \mid A_0, O_0] = \mathbb{E}[\mathbb{E}[b'_W(A_0, O_0^-) \mid A_0, O_0, S_0] \mid A_0, O_0] \\
&= \mathbb{E}[\mathbb{E}[b'_W(A_0, O_0^-) \mid A_0, S_0] \mid A_0, O_0] && (O_0 \perp O_0^- \mid A_0, S_0) \\
&= \mathbb{E}[1/\pi^b(A_0 \mid S_0) \mid A_0, O_0] && \text{(Definition of bridge functions)} \\
&= \mathbb{E}[1/\pi^b(A_0 \mid S_0, O_0) \mid A_0, O_0] && (A_0 \perp O_0 \mid S_0) \\
&= 1/P_{\pi_b}(A_0 \mid O_0).
\end{aligned}$$

□

Proof of Lemma 3. First, note

$$\mathbb{E}[\sum_a f(O_0, a) - b_W(A_0, O_0^-) f(O_0, A_0)] = 0, \quad \forall f : \mathcal{O} \times \mathcal{A} \to \mathbb{R},$$

is equivalent to

$$\mathbb{E}[f(O_0, a) - b_W(A_0, O_0^-) \mathrm{I}(A_0 = a) f(O_0, a)] = 0, \quad \forall f : \mathcal{O} \times \mathcal{A} \to \mathbb{R}, \forall a \in \mathcal{A}.$$

This condition is equivalent to

$$\mathbb{E}[f(O_0, a)\{1 - b_W(A_0, O_0^-)\pi^b(a_0 \mid O_0)\}] = 0, \quad \forall f : \mathcal{O} \times \mathcal{A} \to \mathbb{R}.$$

This is equivalent to $b_W(A_0, O_0^-)\pi^b(a_0 \mid O_0) = 1$.

□

Proof of Theorem 1. We prove the importance sampling formula. Let b'_V be a reward bridge function and b_W be an learnable bridge function. Then,

$$\begin{aligned}
J(\pi^e) &= \mathbb{E}[\sum_a b'_V(a, O_0)] = \mathbb{E}[1/P_{\pi^b}(A_0 \mid O_0) b'_V(A_0, O_0)] && \text{(Result of Lemma 1)} \\
&= \mathbb{E}[\mathbb{E}[b_W(A_0, O_0^-) \mid A_0, O_0] b'_V(A_0, O_0)] && \text{(Definition)} \\
&= \mathbb{E}[b_W(A_0, O_0^-) b'_V(A_0, O_0)] \\
&= \mathbb{E}[b_W(A_0, O_0^-) \mathbb{E}[b'_V(A_0, O_0) \mid A_0, S_0]] \\
&= \mathbb{E}[b_W(A_0, O_0^-) \mathbb{E}[R_0 \pi^e(A_0 \mid O_0) \mid A_0, S_0]] && \text{(Definition)} \\
&= \mathbb{E}[b_W(A_0, O_0^-) R_0 \pi^e(A_0 \mid O_0)].
\end{aligned}$$

We prove the value function based formula. Let b'_W be a weight bridge function and b_V be an learnable weight bridge function. Then,

$$\begin{aligned}
J(\pi^e) &= \mathbb{E}[b'_W(A_0, O_0^-) R_0 \pi^e(A_0 \mid O_0)] = \mathbb{E}[b'_W(A_0, O_0^-) \mathbb{E}[R_0 \pi^e(A_0 \mid O_0) \mid A_0, O_0^-]] && \text{(Result of Lemma 1)} \\
&= \mathbb{E}[b'_W(A_0, O_0^-) b_V(A_0, O_0)] && \text{(Definition)} \\
&= \mathbb{E}[\mathbb{E}[b'_W(A_0, O_0^-) \mid A_0, S_0] b_V(A_0, O_0)] \\
&= \mathbb{E}[1/\pi^b(A_0 \mid S_0) b_V(A_0, O_0)] && \text{(Definition)} \\
&= \mathbb{E}[\sum_a b_V(a, O_0)].
\end{aligned}$$

□

Proof of Theorem 2.

$$\begin{aligned} J(\pi^e) &= \mathbb{E}[(1/P_{\pi_b}(A_0 \mid S_0) R_0 \pi^e(A_0 \mid O_0)] \\ &= \mathbb{E}[(1/P_{\pi_b}(A_0 \mid S_0) \mathbb{E}[R_0 \pi^e(A_0 \mid O_0) \mid A_0, S_0)]] \\ &= \mathbb{E}[(1/P_{\pi_b}(A_0 \mid S_0) \mathbb{E}[b_V(A_0, O_0) \mid A_0, S_0)]] && \text{(Completeness)} \\ &= \mathbb{E}[\sum_a b_V(a, O_0)]. \end{aligned}$$

□

E.1.1 Discrete setting

Tennenholtz et al. (2020) assume the following.

Assumption 8. $\Pr(\mathbf{O}_0 \mid A_0 = a, \mathbf{O}_0^-), \Pr(\mathbf{S}_0 \mid A_0 = a, \mathbf{O}_0^-)$ *are invertible.*

This implies $|\mathbf{S}| = |\mathbf{O}|$, which is very strong. Remark this is also assumed in the paper which proposed negative controls Miao et al. (2018). Instead, we require the following weaker assumption:

Assumption 9. $\operatorname{rank}(\Pr(\mathbf{O}_0 \mid \mathbf{S}_0)) = |\mathcal{S}|$ *and* $\operatorname{rank}(\mathbf{S}_0 \mid \mathbf{O}_0, A_0 = a)) = |\mathcal{S}|$.

The first assumption ensure the existence of weight bridge functions. The second assumption ensures the completeness assumption. We show this is equivalent to the assumption in Nair and Jiang (2021):

Assumption 10. $\operatorname{rank}(\Pr(\mathbf{O}_0 \mid A_0 = a, \mathbf{O}_0^-)) = |\mathcal{S}|$.

Proof of Lemma 5. Assumption 10 implies Assumption 9 since

$$|\mathcal{S}| = \operatorname{rank}(\Pr(\mathbf{O}_0 \mid A_0 = a, \mathbf{O}_0^-)) \leq \min(\operatorname{rank}(\Pr(\mathbf{O}_0 \mid A_0 = a, \mathbf{S}_0)), \operatorname{rank}(\Pr(\mathbf{S}_0 \mid A_0 = a, \mathbf{O}_0^-))).$$

Then, Assumption 10 is concluded.

Assumption 9 implies assumption 10 from Sylvester's rank inequality:

$$\begin{aligned} |\mathcal{S}| &= \operatorname{rank}(\Pr(\mathbf{O}_0 \mid A_0 = a, \mathbf{S}_0)) && \text{(From Assumption 9)} \\ &= \operatorname{rank}(\Pr(\mathbf{O}_0 \mid A_0 = a, \mathbf{S}_0)) + \operatorname{rank}(\Pr(\mathbf{S}_0 \mid A_0 = a, \mathbf{O}_0^-)) - |\mathcal{S}| && \text{(See argument above)} \\ &\leq \operatorname{rank}(\Pr(\mathbf{O}_0 \mid A_0 = a, \mathbf{S}_0) \Pr(\mathbf{S}_0 \mid A_0 = a, \mathbf{O}_0^-)) && \text{(Sylvester's rank inequality:)} \\ &= \operatorname{rank}(\Pr(\mathbf{O}_0 \mid A_0 = a, \mathbf{O}_0^-)). \end{aligned}$$

Besides,

$$\operatorname{rank}(\Pr(\mathbf{O}_0 \mid A_0 = a, \mathbf{O}_0^-)) \leq \min(\operatorname{rank}(\Pr(\mathbf{O}_0 \mid A_0 = a, \mathbf{S}_0)), \operatorname{rank}(\Pr(\mathbf{S}_0 \mid A_0 = a, \mathbf{O}_0^-))) = |\mathcal{S}|.$$

Thus,

$$\operatorname{rank}(\Pr(\mathbf{O}_0 \mid A_0 = a, \mathbf{O}_0^-)) = |\mathcal{S}|.$$

□

We finally remark the existence of weight bridge functions is equal to the one

Assumption 11. *Suppose* $P_{\pi^b}(O_0^- = o^- \mid A_0 = a) > 0, P_{\pi^b}(S_0 = s \mid A_0 = a) > 0$ *for any* $o^- \in \mathcal{O}, s \in \mathcal{S}, a \in \mathcal{A}$. *Then,* $\operatorname{rank}(\Pr(\mathbf{O}_0 \mid \mathbf{S}_0)) = |\mathcal{S}|$ *and* $\operatorname{rank}(\Pr(\mathbf{S}_0 \mid A_0 = a.\mathbf{O}_0^-)) = |\mathcal{S}|$ *are equivalent.*

Proof. We have

$$\{\Pr(\mathbf{S}_0 \mid A_0 = a, \mathbf{O}_0^-)\}^\top = \Lambda_1 \Pr(\mathbf{O}_0^- \mid A_0 = a, \mathbf{S}_0)\Lambda_2$$

where Λ_1 is a $|\mathcal{O}| \times |\mathcal{O}|$ matrix s.t. (i,i)-th element is $1/P_{\pi^b}(O_0^- = o^- \mid A_0 = a)$ and Λ_2 is a $|\mathcal{S}| \times |\mathcal{S}|$ matrix s.t. (i,i)-th element is $P_{\pi^b}(S_0 = s \mid A_0 = a)$. Thus, $\mathrm{rank}(\Pr(\mathbf{O}_0^- \mid \mathbf{S}_0)) = \mathrm{rank}(\Pr(\mathbf{O}_0^- \mid A_0 = a, \mathbf{S}_0)) = \mathrm{rank}(\Pr(\mathbf{S}_0 \mid A_0 = a, \mathbf{O}_0^-)) \geq |\mathcal{S}|$. □

Finally, we give the identification formula:

Proof of Theorem 15. We have

$$\Pr(\mathbf{O}_0 \mid \mathbf{O}_0^-, a) = \Pr(\mathbf{O}_0 \mid \mathbf{S}_0, a)\Pr(\mathbf{S}_0 \mid \mathbf{O}_0^-, a), \quad \Pr(r, o \mid \mathbf{O}_0^-, a) = \Pr(r, o \mid \mathbf{S}_0, a)\Pr(\mathbf{S}_0 \mid \mathbf{O}_0^-, a).$$

Let the SVD decomoposition of $\Pr(\mathbf{O}_0 \mid \mathbf{O}_0^-, a)$ be ABC, where A is a $|\mathcal{O}| \times |\mathcal{S}|$ matrix, B is a $|\mathcal{S}| \times |\mathcal{S}|$ matrix, and C is a $|\mathcal{S}| \times |\mathcal{O}|$ matrix.

$$\mathrm{I} = A^\top \Pr(\mathbf{O}_0 \mid \mathbf{O}_0^-, a)C^\top B^{-1} = A^\top \Pr(\mathbf{O}_0 \mid \mathbf{S}_0, a)\Pr(\mathbf{S}_0 \mid \mathbf{O}_0^-, a)C^\top B^{-1}.$$

Therefore, we have

$$A^\top \Pr(\mathbf{O}_0 \mid \mathbf{S}_0) = \{\Pr(\mathbf{S}_0 \mid \mathbf{O}_0^-, a)C^\top B^{-1}\}^{-1}.$$

noting $\Pr(\mathbf{S}_0 \mid \mathbf{O}_0^-, a)C^\top B^{-1}$ is full-rank matrix from the assumption.

In addition, we have

$$\Pr(r, o \mid \mathbf{O}_0^-, a)C^\top B^{-1} = \Pr(r, o \mid \mathbf{S}_0, a)\Pr(\mathbf{S}_0 \mid \mathbf{O}_0^-, a)C^\top B^{-1}. \tag{14}$$

Thus,

$$\begin{aligned}
\Pr(r, o \mid \mathbf{S}_0, a) &= \Pr(r, o \mid \mathbf{O}_0^-, a)C^\top B^{-1}\{\Pr(\mathbf{S}_0 \mid \mathbf{O}_0^-, a)C^\top B^{-1}\}^{-1} && \text{(From (14))} \\
&= \Pr(r, o \mid \mathbf{O}_0^-, a)C^\top B^{-1}A^\top \Pr(\mathbf{O}_0 \mid \mathbf{S}_0, a) \\
&= \Pr(r, o \mid \mathbf{O}_0^-, a)\{\Pr(\mathbf{O}_0 \mid \mathbf{O}_0^-, a)\}^+ \Pr(\mathbf{O}_0 \mid \mathbf{S}_0).
\end{aligned}$$

Besides,

$$\begin{aligned}
J &= \sum_{a,r,o,s} rp(r, o \mid s, a)\pi^e(a \mid o)p(s) && \text{(Definition)} \\
&= \sum_{a,r,o,s} \Pr(r, o \mid \mathbf{O}_0^-, a)\{\Pr(\mathbf{O}_0 \mid \mathbf{O}_0^-, a)\}^+ \Pr(\mathbf{O}_0 \mid s)\pi^e(A \mid O)p(s) \\
&= \sum_{a,r,o} r\pi^e(a \mid o)\Pr(r, o \mid \mathbf{O}_0^-, a)\{\Pr(\mathbf{O}_0 \mid \mathbf{O}_0^-, a)\}^+ \Pr(\mathbf{O}_0).
\end{aligned}$$

□

E.2 Proof of Section 4.1

We define $Q(s,a) \coloneqq \mathbb{E}[b'_V(a,O) \mid A=a, S=s]$. We show this $Q(s,a)$ satisfies a recursion formula.

Lemma 6.

$$Q(S,A) = \mathbb{E}[R\pi^e(A \mid O) \mid A,S] + \gamma \mathbb{E}\left[\sum_{a'} Q(S^+,a')\pi^e(A \mid O) \mid A,S\right].$$

Proof of Lemma 6. From the definition,

$$Q(s,a) = \mathbb{E}[R_0\pi^e(a \mid O_0) \mid A_0=a, S_0=s] + \mathbb{E}_{\pi^e}\left[\sum_{t=1}^{\infty} \gamma^t R_t \pi^e(a \mid O_0) \mid A_0=a, S_0=s\right].$$

Then, we have

$$\begin{aligned}
&\mathbb{E}_{\pi^e}\left[\sum_{t=1}^{\infty} \gamma^t R_t \pi^e(a \mid O_0) \mid A_0=a, S_0=s\right] \\
&= \gamma \mathbb{E}\left[\sum_{a'} \mathbb{E}_{\pi^e}[\sum_{t=1}^{\infty} \gamma^{t-1} R_t \mid S_1, A_1=a']\pi^e(a' \mid O_1)\pi^e(a \mid O_0) \mid A_0=a, S_0=s\right] \\
&= \gamma \mathbb{E}\left[\sum_{a'} \mathbb{E}_{\pi^e}[\sum_{t=1}^{\infty} \gamma^{t-1} R_t \mid S_1, A_1=a']\mathbb{E}[\pi^e(a' \mid O_1) \mid S_1, A_1=a']\pi^e(a \mid O_0) \mid A_0=a, S_0=s\right] \\
&= \gamma \mathbb{E}\left[\sum_{a'} \mathbb{E}_{\pi^e}[\sum_{t=1}^{\infty} \gamma^{t-1} R_t \pi^e(a' \mid O_1) \mid S_1, A_1=a']\pi^e(a \mid O_0) \mid A_0=a, S_0=s\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (R_1, \cdots, R_{H-1} \perp O_1 \mid S_1, A_1) \\
&= \gamma \mathbb{E}\left[\sum_{a'} Q^{\pi^e}(a', S_1)\pi^e(a \mid O_0) \mid A_0=a, S_0=s\right].
\end{aligned}$$

In conclusion,

$$Q(S,A) = \mathbb{E}[R\pi^e(a \mid O) \mid A,S] + \gamma \mathbb{E}\left[\sum_{a'} Q(S^+,a')\pi^e(A \mid O) \mid A,S\right].$$

□

By using the above lemma, we show learnable reward bridge functions are reward bridge functions. This is a part of the statement in Theorem 3.

Lemma 7 (Learnable reward bridge functions are reward bridge functions)**.**

$$\mathbb{E}[b'_V(A,O) \mid A,O^-] = \mathbb{E}[R\pi^e(a \mid O) \mid A,O^-] + \gamma \mathbb{E}\left[\sum_{a'} b'_V(a',O^+)\pi^e(A \mid O) \mid A,O^-\right].$$

Proof. From Lemma 6, we have

$$\begin{aligned}
\mathbb{E}[b'_V(A,O) \mid S,A] &= \mathbb{E}[R\pi^e(a \mid O) \mid A,S] + \gamma \mathbb{E}\left[\sum_{a'} \mathbb{E}[b'_V(a',O^+)|S^+]\pi^e(A \mid O) \mid A,S\right] \\
&= \mathbb{E}[R\pi^e(a \mid O) \mid A,S] + \gamma \mathbb{E}\left[\sum_{a'} b'_V(a',O^+)\pi^e(A \mid O) \mid A,S\right].
\end{aligned}$$

Then, from $O, R, O^+ \perp O^- \mid S, A$, we have

$$\mathbb{E}[\mathbb{E}[b'_V(A,O) \mid S,A]|A,O^-] = \mathbb{E}[\mathbb{E}[b'_V(A,O) \mid S,A,O^-]|A,O^-] = \mathbb{E}[b'_V(A,O) \mid A,O^-]$$
$$\mathbb{E}[\mathbb{E}[R\pi^e(A \mid O) \mid A,S]|A,O^-] = \mathbb{E}[\mathbb{E}[R\pi^e(A \mid O) \mid A,S,O^-]|A,O^-] = \mathbb{E}[R\pi^e(A \mid O) \mid A,O^-]$$
$$\mathbb{E}\left[\left[\sum_{a'} b'_V(a',O^+)\pi^e(A \mid O) \mid A,S\right] \mid A,O^-\right] = \mathbb{E}\left[\sum_{a'} b'_V(a',O^+)\pi^e(A \mid O) \mid A,O^-\right].$$

In conclusion,

$$\mathbb{E}[b'_V(A,O) \mid A,O^-] = \mathbb{E}[R\pi^e(A \mid O) \mid A,O^-] + \gamma\mathbb{E}\left[\sum_{a'} b'_V(a',O^+)\pi^e(A \mid O) \mid A,O^-\right].$$

□

Next, we show learnable weight bridge functions are action bridge functions. This is a part of the statement in Theorem 3. Here, letting $w(s) = d_{\pi^e}(s)/b(s)$, recall

$$\mathbb{E}\left[\gamma w(S)\frac{\pi^e(A|O)}{\pi^b(A|S)}f(S') - w(S)f(S)\right] + (1-\gamma)\mathbb{E}_{\tilde{S}\sim\nu}[f(\tilde{S})] = 0,\ \forall f : \mathcal{S} \to \mathbb{R} \tag{15}$$

from Kallus and Uehara (2019, Lemma 16) which is a modification of the lemma in Liu et al. (2018). Recall ν is an initial distribution.

Lemma 8 (Learnable weight bridge functions are weight bridge functions.)**.**

$$\mathbb{E}\left[\gamma b'_W(A,O^-)\pi^e(A|O)f(O^+,a') - \mathrm{I}(A=a')b'_W(A,O^-)f(O,A)\right] + (1-\gamma)\mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[f(\tilde{O},a')] = 0,\ \forall f : \mathcal{S} \to \mathbb{R}.$$

Similarly,

$$\mathbb{E}\left[\sum_{a'\in\mathcal{A}} \gamma b'_W(A,O^-)\pi^e(A|O)f(O^+,a') - b'_W(A,O^-)f(O,A)\right] + (1-\gamma)\mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_{a'\in\mathcal{A}} f(\tilde{O},a')] = 0,\ \forall f : \mathcal{S} \to \mathbb{R}.$$

Proof. We first prove the first statement:

$$\begin{aligned}
&\mathbb{E}\left[\gamma b'_W(A,O^-)\pi^e(A \mid O)f(O^+,a') - \mathrm{I}(A=a')b'_W(A,O^-)f(O,A)\right] \\
&= \mathbb{E}\left[\gamma\mathbb{E}[b'_W(A,O^-) \mid A,S,O^+,O]\pi^e(A \mid O)f(O^+,a') - \mathrm{I}(A=a')\mathbb{E}[b'_W(A,O^-) \mid A,S,O]f(O,A)\right] \\
&= \mathbb{E}\left[\gamma\mathbb{E}[b'_W(A,O^-) \mid A,S]\pi^e(A \mid O)f(O^+,a') - \mathrm{I}(A=a')\mathbb{E}[b'_W(A,O^-) \mid A,S]f(O,A)\right] \\
&= \mathbb{E}\left[\gamma w(S)/\pi^b(A \mid S)\pi^e(A|O)f(O^+,a') - w(S)\mathrm{I}(A=a')/\pi^b(A \mid S)f(O,A)\right] && \text{(Definition)} \\
&= \mathbb{E}\left[\gamma w(S)/\pi^b(A \mid S)\pi^e(A|O)\mathbb{E}[f(O^+,a') \mid S^+,A,O,S] - w(S)\mathrm{I}(A=a')/\pi^b(A \mid S)\mathbb{E}[f(O,A) \mid S,A=a']\right] \\
&= \mathbb{E}\left[\gamma w(S)\pi^e(A|O)/\pi^b(A \mid S)\mathbb{E}[f(O^+,a') \mid S^+] - w(S)\mathbb{E}[f(O,a') \mid S]\right] \\
&= -(1-\gamma)\mathbb{E}_{S\sim\nu}[\mathbb{E}[f(\tilde{O},a') \mid S]] && \text{(From (15))} \\
&= -(1-\gamma)\mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[f(\tilde{O},a')].
\end{aligned}$$

The second statement is proved by taking summation over the action space. □

Finally, we show if bridge functions exist, we can identify the policy value. Before proceeding, we prove the following helpful lemma.

Lemma 9 (Identification formula with (unlearnable) bridge functions)**.** *Assume the existence of bridge functions. Then, let b'_V, b'_W be any bridge functions. Then, we have*

$$J(\pi^e) = \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_a b'_V(a,\tilde{O})], \quad J(\pi^e) = (1-\gamma)^{-1}\mathbb{E}[b'_W(A,O^-)R\pi^e(A\mid O)].$$

Proof. The first formula is proved as follows:

$$\begin{aligned}
&\mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_a b'_V(a,\tilde{O})] = \mathbb{E}_{S\sim\nu}[\mathbb{E}[\sum_a b'_V(a,\tilde{O})\mid S]] = \mathbb{E}_{S\sim\nu}[\mathbb{E}[\sum_a b'_V(a,\tilde{O})\mid A=a,S]] \\
&= \mathbb{E}_{S_0\sim\nu}[\mathbb{E}[\sum_a b'_V(a,\tilde{O}_0)\mid A_0=a,S_0]] \\
&= \mathbb{E}_{S_0\sim\nu}[\sum_a \mathbb{E}_{\pi^e}[\sum_{t=0}^{\infty}\gamma^t R_t\pi^e(a\mid O_0)|A_0=a,S_0]] && \text{(Definition)} \\
&= \mathbb{E}_{S_0\sim\nu}[\sum_a \mathbb{E}_{\pi^e}[\mathbb{E}_{\pi^e}[\sum_{t=0}^{\infty}\gamma^t R_t\mid A_0=a,S_0,O_0]\pi^e(a\mid O_0)|A_0=a,S_0]] \\
&= \mathbb{E}_{S_0\sim\nu}[\sum_a q^{\pi^e}(S_0,a)\mathbb{E}_{\pi^e}[\pi^e(a\mid O_0)|A_0=a,S_0]] \\
&= \mathbb{E}_{S_0\sim\nu}[\sum_a q^{\pi^e}(S_0,a)\mathbb{E}_{\pi^e}[\pi^e(a\mid O_0)|S_0]] = \mathbb{E}_{S_0\sim\nu}[\sum_a q^{\pi^e}(S_0,a)\pi^e(a\mid O_0)] \\
&= J(\pi^e).
\end{aligned}$$

Here, we define $\mathbb{E}_{\pi^e}[\sum_{t=0}^{\infty}\gamma^t R_t|A_0=a,S_0=s] = q^{\pi^e}(a,s)$.

The second formula is proved as follows:

$$\begin{aligned}
\mathbb{E}[b'_W(A,O^-)R\pi^e(A\mid O)] &= \mathbb{E}[\mathbb{E}[b'_W(A,O^-)\mid A,S,R,O]\pi^e(A\mid O)R] \\
&= \mathbb{E}[\mathbb{E}[b'_W(A,O^-)\mid A,S]\pi^e(A\mid O)R] && (O^-\perp R,O\mid S,A) \\
&= \mathbb{E}\left[\frac{w(S)}{\pi^b(A|S)}\pi^e(A\mid O)R\right] && \text{(Definition)} \\
&= J(\pi^e).
\end{aligned}$$

□

We are ready to give the proof of the final identification formula Theorem 3. The statement is as follows.

Theorem 17. *Assume the existence of bridge functions. Let b_V, b_W be any learnable bridge functions. Then,*

$$J(\pi^e) = \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_{a'} b_V(\tilde{O},a')], \quad J(\pi^e) = (1-\gamma)^{-1}\mathbb{E}[b_W(A,O^-)R\pi^e(A\mid O)].$$

Proof. First, we prove the first formula. This is concluded by

$$\begin{aligned}
J(\pi^e) &= (1-\gamma)^{-1}\mathbb{E}[b'_W(A,O^-)R\pi^e(A\mid O)] && \text{(From Lemma 9)} \\
&= (1-\gamma)^{-1}\mathbb{E}[b'_W(A,O^-)\mathbb{E}[R\pi^e(A\mid O)\mid A,O^-]] \\
&= (1-\gamma)^{-1}\mathbb{E}\left[b'_W(A,O^-)\mathbb{E}\left[\gamma\sum_{a'} b_V(a',O^+)\pi^e(A\mid O) - b_V(A,O)\mid A,O^-\right]\right] \\
& && (b'_W \text{ is also an learnable action bridge function.}) \\
&= \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_{a'} b_V(\tilde{O},a')]. && \text{(From Lemma 8)}
\end{aligned}$$

Next, we prove the second formula. This is concluded by

$$\begin{aligned}
J(\pi^e) &= \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_{a'} b'_V(\tilde{O}, a')] && \text{(From Lemma 9)} \\
&= (1-\gamma)^{-1}\mathbb{E}\left[b_W(A, O^-)\mathbb{E}\left[\gamma\sum_{a'} b'_V(a', O^+)\pi^e(A \mid O) - b'_V(A, O) \mid A, O^-\right]\right] && \text{(Definition of } b_W\text{)} \\
&= (1-\gamma)^{-1}\mathbb{E}[b_W(A, O^-)\mathbb{E}[R\pi^e(A \mid O) \mid A, O^-]] && (b'_V \text{ is also an learnable reward function.}) \\
&= (1-\gamma)^{-1}\mathbb{E}[b_W(A, O^-)R\pi^e(A \mid O)].
\end{aligned}$$

□

Proof of Theorem 4. By the assumption $O, R, O^+ \perp O^- \mid S, A$ we see any b_V satisfies

$$\mathbb{E}\left[\mathbb{E}\left[\gamma\sum_{a'} b_V(a', O^+) + R\pi^e(A \mid O) - b_V(A, O) \mid S, A\right] \mid A, O^-\right] = 0.$$

From the completeness assumption,

$$\mathbb{E}\left[\gamma\sum_{a'} b_V(a', O^+) + R\pi^e(A \mid O) - b_V(A, O) \mid S, A\right] = 0. \tag{16}$$

Thus, we have

$$\begin{aligned}
J(\pi^e) &= (1-\gamma)^{-1}\mathbb{E}[w(S)\pi^e(A \mid O)/P_{\pi_b}(A \mid S)R] \\
&= (1-\gamma)^{-1}\mathbb{E}[w(S)/P_{\pi_b}(A \mid S)\mathbb{E}[R\pi^e(A \mid O) \mid A, S]] \\
&= (1-\gamma)^{-1}\mathbb{E}\left[w(S)/P_{\pi_b}(A \mid S)\mathbb{E}\left[\gamma\sum_{a'} b_V(a', O^+) + R\pi^e(A \mid O) - b_V(A, O) \mid S, A\right]\right] + \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_{a'} b_V(\tilde{O}, a')] \\
&= \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_{a'} b_V(\tilde{O}, a')]. && \text{(From (16))}
\end{aligned}$$

□

E.3 Proof of Section 4.3.1

Proof of Theorem 5. By simple algebra, the estimator is written as

$$\hat{b}_V = \inf_{f\in\mathcal{V}}\sup_{g\in\mathcal{V}^\dagger} \mathbb{E}_{\mathcal{D}}[(\mathcal{Z}f)^2] - \mathbb{E}_{\mathcal{D}}[(\mathcal{Z}f - g(A, O^-))^2]$$

where $\mathcal{Z}f = R\pi^e(A \mid O) + \sum_{a'} f(a', O^+)\pi^e(A \mid O)$. In this proof, we define

$$\bar{C} = \max(C_{\mathcal{V}}, C_{\mathcal{V}^\dagger}, 1).$$

Furthermore, c is some universal constant.

Show the convergence of inner maximizer For fixed $f \in \mathcal{V}$, we define

$$\hat{g}_f = \sup_{g \in \mathcal{V}^\dagger} -\mathbb{E}_{\mathcal{D}}[(\mathcal{Z}f - g(A, O^-))^2],$$

we want to prove with probability $1-\delta$:

$$\forall f \in \mathcal{V} : |\mathbb{E}_{\mathcal{D}}[(\mathcal{Z}f - \hat{g}_f(A, O^-))^2] - \mathbb{E}_{\mathcal{D}}[(\mathcal{Z}f - \mathcal{T}f)^2(A, O^-)]| \leq \frac{c \log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)}{n}.$$

Note since $\mathcal{B}f$ is the Bayes optimal regressor, we have

$$\mathbb{E}[(\mathcal{Z}f - \mathcal{T}f(A, O^-))^2 - (\mathcal{Z}f - g(A, O^-))^2] = \mathbb{E}[(\mathcal{T}f - g(A, O^-))^2]. \tag{17}$$

Hereafter, to simplify the notation, we often drop (A, O^-). Then, from Bernstein's inequality, with probability $1-\delta$,

$$\forall f \in \mathcal{V}, \forall g \in \mathcal{V}^\dagger : |\{\mathbb{E}_{\mathcal{D}} - \mathbb{E}\}[(\mathcal{Z}f - \mathcal{T}f)^2 - (\mathcal{Z}f - g)^2]| \leq c\bar{C}\sqrt{\frac{\mathbb{E}[(\mathcal{T}f - g)^2] \log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)}{n}} + \frac{c\bar{C}^2 \log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)}{n}. \tag{18}$$

Hereafter, we condition on the above event. In addition, from Bellman closedness assumption $\mathcal{T}\mathcal{V} \subset \mathcal{V}^\dagger$,

$$\mathbb{E}_{\mathcal{D}}[(\mathcal{Z}f - \hat{g}_f)^2 - (\mathcal{Z}f - \mathcal{T}f)^2] \leq 0.$$

Then,

$$\begin{aligned}
\mathbb{E}[(\mathcal{T}f - \hat{g}_f)^2] &= \mathbb{E}[(\mathcal{Z}f - \mathcal{T}f)^2 - (\mathcal{Z}f - \hat{g}_f)^2] \\
&\leq |\{\mathbb{E}_{\mathcal{D}} - \mathbb{E}\}[(\mathcal{Z}f - \mathcal{T}f)^2 - (\mathcal{Z}f - \hat{g}_f)^2]| + \mathbb{E}_{\mathcal{D}}[(\mathcal{Z}f - \hat{g}_f)^2 - (\mathcal{Z}f - \mathcal{B}f)^2] \\
&\leq |\{\mathbb{E}_{\mathcal{D}} - \mathbb{E}\}[(\mathcal{Z}f - \mathcal{T}f)^2 - (\mathcal{Z}f - \hat{g}_f)^2]| \\
&\leq c\bar{C}\sqrt{\frac{\mathbb{E}[(\mathcal{T}f - \hat{g}_f)^2] \log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)}{n}} + \frac{c\bar{C}^2 \log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)}{n}. && \text{(From (18))}
\end{aligned}$$

This concludes

$$\mathbb{E}[(\mathcal{T}f - \hat{g}_f)^2] \leq c\text{Error}, \quad \text{Error} := \frac{\bar{C}^2 \log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)}{n}.$$

Then,

$$\begin{aligned}
|\mathbb{E}_{\mathcal{D}}[(\mathcal{Z}f - \hat{g}_f)^2] - \mathbb{E}_{\mathcal{D}}[(\mathcal{Z}f - \mathcal{T}f)^2]| &\leq \mathbb{E}[(\mathcal{Z}f - \mathcal{T}f)^2 - (\mathcal{Z}f - \hat{g}_f)^2] + c\text{Error} && \text{(From (18))} \\
&= \mathbb{E}[(\mathcal{T}f - \hat{g}_f)^2] + c\text{Error} && \text{(From (17))} \\
&\leq 2c\text{Error}.
\end{aligned}$$

Show the convergence of outer minimizer From the first observation, we can see

$$\mathbb{E}_{\mathcal{D}}[(\mathcal{Z}\hat{b}_V)^2 - (\mathcal{Z}\hat{b}_V - \mathcal{T}\hat{b}_V(A, O^-))^2] \leq \mathbb{E}_{\mathcal{D}}[(\mathcal{Z}b_V)^2 - (\mathcal{Z}b_V - \mathcal{T}b_V(A, O^-))^2] + c\text{Error}.$$

Note

$$\mathbb{E}[(\mathcal{Z}f)^2 - (\mathcal{Z}f - \mathcal{T}f(A, O^-))^2 - \{(\mathcal{Z}b_V)^2 - (\mathcal{Z}b_V - \mathcal{T}b_V(A, O^-))^2\}] = \mathbb{E}[(\mathcal{T}f)(A, O^-)^2]$$

Here, from Bernstein's inequality, with probability $1-\delta$,

$$\begin{aligned}\forall f \in \mathcal{V}, &(\mathbb{E}_{\mathcal{D}} - \mathbb{E})[(\mathcal{Z}f)^2 - (\mathcal{Z}f - \mathcal{B}f)^2 - \{(\mathcal{Z}b_V)^2 - (\mathcal{Z}b_V - \mathcal{B}b_V)^2\}] \\ &\leq \bar{C}\sqrt{\mathbb{E}[(\mathcal{T}f)^2(A, O^-)]\frac{\log(|\mathcal{V}|/\delta)}{n}} + \frac{\bar{C}^2\log(|\mathcal{V}|/\delta)}{n}.\end{aligned}$$

Hereafter, we condition on the above event. Then,

$$\begin{aligned}\mathbb{E}[(\mathcal{B}\hat{b}_V - \hat{b}_V)^2] &= \mathbb{E}[(\mathcal{Z}\hat{b}_V)^2 - (\mathcal{Z}\hat{b}_V - \mathcal{T}\hat{b}_V)^2 - \{(\mathcal{Z}b_V)^2 - (\mathcal{Z}b_V - \mathcal{T}b_V)^2\}] \\ &\leq (\mathbb{E}_{\mathcal{D}} - \mathbb{E})[(\mathcal{Z}\hat{b}_V)^2 - (\mathcal{Z}\hat{b}_V - \mathcal{T}\hat{b}_V)^2 - \{(\mathcal{Z}b_V)^2 - (\mathcal{Z}b_V - \mathcal{T}b_V)^2\}] \\ &+ \mathbb{E}_{\mathcal{D}}[(\mathcal{Z}\hat{b}_V)^2 - (\mathcal{Z}\hat{b}_V - \mathcal{B}\hat{b}_V)^2 - \{(\mathcal{Z}b_V)^2 - (\mathcal{Z}b_V - \mathcal{T}b_V)^2\}] \\ &\leq \bar{C}\sqrt{\mathbb{E}[(\mathcal{T}\hat{b}_V)^2(A, O^-)]\frac{\log(|\mathcal{V}|c/\delta)}{n}} + \frac{\bar{C}^2\log(|\mathcal{V}|c/\delta)}{n} + c\mathrm{Error}.\end{aligned}$$

This concludes

$$\mathbb{E}[(\mathcal{T}\hat{b}_V)^2(A, O^-)] \leq c\mathrm{Error}.$$

□

Proof of Theorem 6.

First Statement Recall

$$J(\pi^e) = \mathbb{E}[b_W(A, O^-)\mathbb{E}[R\pi^e(A \mid O) \mid A, O^-]$$

from Theorem 17 and the existence of b_W. Furthermore,

$$\begin{aligned}&(1-\gamma)\mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_{a'}\hat{b}_V(\tilde{O}, a')] \\ &= \mathbb{E}\left[b_W(A, O^-)\mathbb{E}\left[\gamma\sum_{a'}\hat{b}_V(a', O^+)\pi^e(A \mid O) - \hat{b}_V(A, O) \mid A, O^-\right]\right].\end{aligned}$$

(b_W's are also observe bridge functions form Theorem 3)

Thus, this concludes

$$\begin{aligned}&J(\pi^e) - \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_{a'}\hat{b}_V(\tilde{O}, a')] \\ &= (1-\gamma)^{-1}\mathbb{E}\left[b_W(A, O^-)\mathbb{E}\left[\gamma\sum_{a'}\hat{b}_V(a', O^+)\pi^e(A \mid O) + R\pi^e(A \mid O) - \hat{b}_V(A, O) \mid A, O^-\right]\right] \\ &= (1-\gamma)^{-1}\mathbb{E}\left[b_W(A, O^-)\mathcal{T}\hat{b}_V(A, O^-)\right].\end{aligned}$$

Then, from CS inequality, we have

$$|J(\pi^e) - \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_{a'}\hat{b}_V(\tilde{O}, a')]| \leq (1-\gamma)^{-1}\mathbb{E}[b_W(A, O^-)^2]^{1/2}\mathbb{E}[\{\mathcal{T}\hat{b}_V(A, O^-)\}^2]^{1/2}.$$

This concludes the final statement. □

Proof of Theorem 7. Recall

$$J(\pi^e) = \mathbb{E}[w(S)/P_{\pi_b}(A \mid S)\mathbb{E}[R\pi^e(A \mid O) \mid A, S]]$$

Furthermore,

$$\begin{aligned}
&(1-\gamma)\mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_{a'}\hat{b}_V(\tilde{O}, a')] \\
&= \mathbb{E}\left[w(S)/P_{\pi_b}(A \mid S)\mathbb{E}\left[\gamma\sum_{a'}\hat{b}_V(a', O^+)\pi^e(A \mid O) - \hat{b}_V(A, O) \mid A, S\right]\right].
\end{aligned}$$

Thus, this concludes

$$\begin{aligned}
&J(\pi^e) - \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_{a'}\hat{b}_V(\tilde{O}, a')] \\
&= (1-\gamma)^{-1}\mathbb{E}\left[w(S)/P_{\pi_b}(A \mid S)\mathbb{E}\left[\gamma\sum_{a'}\hat{b}_V(a', O^+)\pi^e(A \mid O) + R\pi^e(A \mid O) - \hat{b}_V(A, O) \mid A, S\right]\right] \\
&= (1-\gamma)^{-1}\mathbb{E}\left[w(S)/P_{\pi_b}(A \mid S)\mathcal{T}_S\hat{b}_V(A, S)\right].
\end{aligned}$$

Then, from CS inequality, we have

$$|J(\pi^e) - \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_{a'}\hat{b}_V(\tilde{O}, a')]| \leq (1-\gamma)^{-1}\mathbb{E}[\{w(S)/P_{\pi_b}(A \mid S)\}^2]^{1/2}\mathbb{E}[\mathcal{T}_S\hat{b}_V(A, S)^2]^{1/2}.$$

Recalling $\mathbb{E}[\{\mathcal{T}_S\hat{b}_V(A, S)\}^2] \leq \kappa\mathbb{E}[\{\mathcal{T}\hat{b}_V(A, O^-)\}^2]$, this concludes the final statement.

□

E.4 Proof of Section 4.3.2

Proof of Theorem 8. The first statement $J(\pi^e) = \mathbb{E}[J(f, b_V)]$ is proved by Lemma 8:

$$\begin{aligned}
\mathbb{E}[J(f, b_V)] &= \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_a b_V(a, \tilde{O})] + \mathbb{E}\left[\frac{f(A, O^-)}{1-\gamma}\{\{R + \gamma\sum_{a'}b_V(a', O^+)\}\pi^e(A \mid O) - b_V(A, O)\}\right] \\
&= \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_a b_V(a, \tilde{O})] && \text{(Definition of reward bridge functions)} \\
&= J(\pi^e). && \text{(From Theorem 17)}
\end{aligned}$$

The second statement $J(\pi^e) = \mathbb{E}[J(b_W, f)]$ is proved by Lemma 7:

$$\begin{aligned}
\mathbb{E}[J(b_W, f)] &= \mathbb{E}_{\tilde{O}\sim\nu_{\mathcal{O}}}[\sum_a f(a, \tilde{O})] + \mathbb{E}\left[\frac{b_W(A, O^-)}{1-\gamma}\{\{R + \gamma\sum_{a'}f(a', O^+)\}\pi^e(A \mid O) - f(A, O)\}\right] \\
&= (1-\gamma)^{-1}\mathbb{E}[R\pi^e(A \mid O)b_W(A, O^-)] && \text{(Definition of weight bridge functions)} \\
&= J(\pi^e). && \text{(From Theorem 17)}
\end{aligned}$$

□

Proof of Theorem 9. Recall we have the observation:

$$Z = \{O^-, A, O, R, O^+\}.$$

We define

$$\tau_k = \prod_{t=1}^{k} \pi_t^e(a_t \mid o_t), \quad \tau_{a:b} = \prod_{t=a}^{b} \pi_t^e(a_t \mid o_t).$$

From Nair and Jiang (2021); Tennenholtz et al. (2020), the target functional is

$$J = \sum_{k=0} \gamma^k \sum_{\mathfrak{j}_{r_k}} r_k \tau_k \Pr(r_k, o_k \mid a_k, \mathbf{O}_{k-1}) \{\Pr(\mathbf{O}_k \mid a_k, \mathbf{O}_{k-1}\}^{-1} \Pr(\mathbf{O}_k, o_{k-1} \mid a_{k-1}, \mathbf{O}_{k-2}) \{\Pr(\mathbf{O}_{k-1} \mid a_{k-1}, \mathbf{O}_{k-2})\}^{-1} \cdots \Pr(\mathbf{O}_0).$$

We use the assumption $\mathrm{rank}(\Pr(\mathbf{O} \mid A = a, \mathbf{O}^-)) = |\mathcal{S}| = |\mathcal{O}|$, i.e., $\Pr(\mathbf{O} \mid A = a, \mathbf{O}^-)$ is full-rank.

We will show the existence of $\phi_{\mathrm{EIF}}(Z)$ s.t.

$$\nabla J(\theta) = \mathbb{E}[\phi_{\mathrm{EIF}}(Z) \nabla \log P(Z)].$$

Then, $\mathbb{E}[\phi_{\mathrm{EIF}}(Z)\phi_{\mathrm{EIF}}^\top(Z)]$ is the Cramér-Rao lower bound.

Before the calculation, we introduce the bridge functions. Due to the assumptions, these are unique. By letting $\mathfrak{j}_k \coloneqq \{o_0, a_0, r_0, a_1, o_1, r_1, \cdots, r_k\}$, we define $b_V(a, o)$ as

$$b_V(a_0, o_0) = \sum_{k=0}^{\infty} \gamma^k \sum_{\mathfrak{j}_{r_k} \setminus (a_0, o_0)} r_k \tau_k \Pr(r_k, o_k \mid a_k, \mathbf{O}_{k-1}) \{\Pr(\mathbf{O}_k \mid a_k, \mathbf{O}_{k-1}\}^{-1} \cdots \{\Pr(\mathbf{O}_0 \mid a_0, \mathbf{O}_{-1})\}^{-1} \mathrm{I}(\mathbf{O}_0 = o_0).$$

Next, we define $b_W(a, o^-)$. Recall

$$d_k^{\pi^e}(o_k) = \sum_{\mathfrak{j}_{o_k} \setminus o_k} \tau_{k-1} \mathrm{I}(\mathbf{O}_k = o_k)^\top \Pr(\mathbf{O}_k, o_{k-1} \mid a_{k-1}, \mathbf{O}_{k-2}) \{\Pr(\mathbf{O}_{k-1} \mid a_{k-1}, \mathbf{O}_{k-2})\}^{-1} \cdots \Pr(\mathbf{O}_0).$$

where $\mathfrak{j}_{o_k} \coloneqq \{o_0, a_0, r_0, a_1, o_1, r_1, \cdots, o_k\}$. Then, we define $b_W(a, o^-)$ as

$$b_W(a, o^-) = \frac{\mathrm{I}(\mathbf{O}^- = o^-)}{\Pr(a, o^-)} \{\Pr(\mathbf{O} \mid a, \mathbf{O}^-\}^{-1} d^{\pi^e}(\mathbf{O}), \quad d^{\pi^e}(o) = \sum_{k=0}^{\infty} \gamma^k d_k^{\pi^e}(o)/(1-\gamma).$$

Now, we are ready to calculate the Cramér-Rao lower bound. We have

$$\nabla J(\theta) = B_1 + B_2 + B_3$$

where

$$B_1 = \sum_{k=0} \gamma^k \sum_{\mathfrak{j}_{r_k}} r_k \tau_k \nabla \Pr(r_k, o_k \mid a_k, \mathbf{O}_{k-1}) \{\Pr(\mathbf{O}_k \mid a_k, \mathbf{O}_{k-1}\}^{-1} \Pr(\mathbf{O}_k, o_{k-1} \mid a_k, \mathbf{O}_{k-2}) \{\Pr(\mathbf{O}_{k-1} \mid a_{k-1}, \mathbf{O}_{k-2})\}^{-1} \cdots \Pr(\mathbf{O}_0),$$

$$B_2 = \sum_{\mathfrak{j}} \sum_{k=0} \gamma^k \sum_{t=0}^{k} r_k \tau_k \Pr(r_k, o_k \mid a_k, \mathbf{O}_{k-1}) \{\Pr(\mathbf{O}_k \mid a_k, \mathbf{O}_{k-1}\}^{-1} \cdots \Pr(\mathbf{O}_{t+1}, o_t \mid a_t, \mathbf{O}_{t-1}) \nabla \{\Pr(\mathbf{O}_t \mid a_t, \mathbf{O}_{t-1})\}^{-1} \cdots \Pr(\mathbf{O}_0),$$

$$B_3 = \sum_{\mathfrak{j}} \sum_{k=0} \gamma^k \sum_{t=0}^{k-1} r_k \tau_k \Pr(r_k, o_k \mid a_k, \mathbf{O}_{k-1}) \{\Pr(\mathbf{O}_k \mid a_k, \mathbf{O}_{k-1}\}^{-1} \cdots \nabla \Pr(\mathbf{O}_{t+1}, o_t \mid a_t, \mathbf{O}_{t-1}) \{\Pr(\mathbf{O}_t \mid a_t, \mathbf{O}_{t-1})\}^{-1} \cdots \Pr(\mathbf{O}_0).$$

Here, we use the notation $\mathfrak{j} = \{s_0, a_0, r_0, \cdots\}$.

We analyze B_1:

$$
\begin{aligned}
B_1 &= \sum_{k=0} \gamma^k \sum_{\mathrm{j}_{r_k}} r_k \tau_k \nabla \Pr(r_k, o_k \mid a_k, \mathbf{O}_{k-1}) \{\Pr(\mathbf{O}_k \mid a_k, \mathbf{O}_{k-1}\}^{-1} \Pr(\mathbf{O}_k, o_{k-1} \mid a_k, \mathbf{O}_{k-2}) \{\Pr(\mathbf{O}_{k-1} \mid a_{k-1}, \mathbf{O}_{k-2})\}^{-1} \cdots \Pr(\mathbf{O}_0) \\
&= (1-\gamma)^{-1} \sum_{r,a,o} r \pi^e(a \mid o) \nabla \Pr(r, o \mid a, \mathbf{O}^-) \{b_W(a, \mathbf{O}^-) \odot \Pr(a, \mathbf{O}^-)\} \\
&= (1-\gamma)^{-1} \sum_{r,a,o,o^-} r \pi^e(a \mid o) \nabla \Pr(r, o \mid a, o^-) b_W(a, o^-) \Pr(a, o^-) \\
&= (1-\gamma)^{-1} \sum_{r,a,o,o^-} r \pi^e(a \mid o) \{\nabla \log \Pr(r, o \mid a, o^-)\} b_W(a, o^-) \Pr(r, o, a, o^-),
\end{aligned}
$$

where $\odot$ is an element-wise product. Then, we have

$$
\begin{aligned}
(1-\gamma) B_1 &= \mathbb{E}[b_W(A, O^-) R \pi^e(A \mid O) \nabla \log \Pr(R, O \mid A, O^-)] \\
&= \mathbb{E}[b_W(A, O^-) \{R \pi^e(A \mid O) - \mathbb{E}[R \pi^e(A \mid O) \mid A, O^-]\} \nabla \log \Pr(R, O \mid A, O^-)] \\
&= \mathbb{E}[b_W(A, O^-) \{R \pi^e(A \mid O) - \mathbb{E}[R \pi^e(A \mid O) \mid A, O^-]\} \nabla \log \Pr(R, O, A, O^-)] \\
&= \mathbb{E}[b_W(A, O^-) \{R \pi^e(A \mid O) - \mathbb{E}[R \pi^e(A \mid O) \mid A, O^-]\} \nabla \log \Pr(O^+, R, O, A, O^-)].
\end{aligned}
$$

Next, we analyze B_2:

$$
\begin{aligned}
&\sum_{\mathrm{j}} \sum_{k=0} \gamma^k \sum_{t=0}^{k} r_k \tau_k \Pr(r_k, o_k \mid a_k, \mathbf{O}_{k-1}) \{\Pr(\mathbf{O}_k \mid a_k, \mathbf{O}_{k-1}\}^{-1} \cdots \Pr(\mathbf{O}_{t+1}, o_t \mid a_t, \mathbf{O}_{t-1}) \nabla \{\Pr(\mathbf{O}_t \mid a_t, \mathbf{O}_{t-1})\}^{-1} \cdots \Pr(\mathbf{O}_0) \\
&= \sum_{\mathrm{j}} \sum_{t=0} \gamma^t \sum_{k=t}^{\infty} \gamma^{k-t} \tau_k r_k \Pr(r_k, o_k \mid a_k, \mathbf{O}_{k-1}) \{\Pr(\mathbf{O}_k \mid a_k, \mathbf{O}_{k-1}\}^{-1} \cdots \Pr(\mathbf{O}_{t+1}, o_t \mid a, \mathbf{O}_{t-1}) \nabla \{\Pr(\mathbf{O}_t \mid a_t, \mathbf{O}_{t-1})\}^{-1} \cdots \Pr(\mathbf{O}_0) \\
&= \sum_{\mathrm{j}} \sum_{t=0} \gamma^t \sum_{k=t}^{\infty} \gamma^{k-t} \tau_k r_k \Pr(r_k, o_k \mid a_k, \mathbf{O}_{k-1}) \{\Pr(\mathbf{O}_k \mid a_k, \mathbf{O}_{k-1}\}^{-1} \cdots \Pr(\mathbf{O}_{t+1}, o_t \mid a, \mathbf{O}_{t-1}) \{\Pr(\mathbf{O}_t \mid a_t, \mathbf{O}_{t-1})\}^{-1} \\
&\times \{\nabla \Pr(\mathbf{O}_t \mid a_t, \mathbf{O}_{t-1})\} \{\Pr(\mathbf{O}_t \mid a_t, \mathbf{O}_{t-1})\}^{-1} \cdots \Pr(\mathbf{O}_0) \\
&= -\sum_{t=0} \gamma^t \tau_t b_V(a_t, \mathbf{O}_t) \{\nabla \Pr(\mathbf{O}_t \mid a_t, \mathbf{O}_{t-1})\} \{\Pr(\mathbf{O}_t \mid a_t, \mathbf{O}_{t-1})\}^{-1} \cdots \Pr(\mathbf{O}_0) \\
&= -(1-\gamma)^{-1} \sum_{a} b_V(a, \mathbf{O}) \{\nabla \Pr(\mathbf{O} \mid a, \mathbf{O}^-) \{b_W(a, \mathbf{O}^-) \odot \Pr(a, \mathbf{O}^-)\}\}.
\end{aligned}
$$

Further, we have

$$
\begin{aligned}
(1-\gamma) B_2 &= \mathbb{E}[b_W(A, O^-) b_V(A, O) \nabla \log \Pr(O \mid A, O^-)] \\
&= \mathbb{E}[b_W(A, O^-) b_V(A, O) \{\nabla \log \Pr(O \mid A, O^-) + \nabla \log \Pr(R, O^+ \mid A, O^-, O)\}] \\
&= \mathbb{E}[b_W(A, O^-) b_V(A, O) \nabla \log \Pr(R, O^+, O \mid A, O^-)] \\
&= \mathbb{E}[b_W(A, O^-) \{b_V(A, O) - \mathbb{E}[b_V(A, O) \mid A, O^-]\} \nabla \log \Pr(R, O^+, O \mid A, O^-)] \\
&= \mathbb{E}[b_W(A, O^-) \{b_V(A, O) - \mathbb{E}[b_V(A, O) \mid A, O^-]\} \nabla \log \Pr(R, O^+, O, A, O^-)].
\end{aligned}
$$

Finally, we analyze B_3:

$$\sum_{\mathrm{j}}\sum_{k=0}\gamma^k\sum_{t=0}^{k-1} r_k\tau_k \Pr(r_k, o_k \mid a_k, \mathbf{O}_{k-1})\{\Pr(\mathbf{O}_k \mid a_k, \mathbf{O}_{k-1}\}^{-1}\cdots\nabla\Pr(\mathbf{O}_{t+1}, o_t \mid a_t, \mathbf{O}_{t-1})\{\Pr(\mathbf{O}_t \mid a_t, \mathbf{O}_{t-1})\}^{-1}\cdots\Pr(\mathbf{O}_0)$$
$$= \sum_{\mathrm{j}}\sum_{t=0}\gamma^t\tau_t\sum_{k=t+1}^{\infty}\gamma^{k-t} r_k\tau_{t+1:k}\Pr(r_k, o_k \mid a_k, \mathbf{O}_{k-1})\{\Pr(\mathbf{O}_k \mid a_k, \mathbf{O}_{k-1}\}^{-1}\cdots\nabla\Pr(\mathbf{O}_{t+1}, o_t \mid a_t, \mathbf{O}_{t-1})\{\Pr(\mathbf{O}_t \mid a_t, \mathbf{O}_{t-1})\}^{-1}\cdots$$
$$= \sum_{\mathrm{j}}\sum_{t=0}\gamma^t\tau_t b_{a_{t+1}}(\mathbf{O}_{t+1})\nabla\Pr(\mathbf{O}_{t+1}, o_t \mid a_t, \mathbf{O}_{t-1})\{\Pr(\mathbf{O}_t \mid a_t, \mathbf{O}_{t-1})\}^{-1}\cdots\Pr(\mathbf{O}_0)$$
$$= (1-\gamma)^{-1}\sum_{\mathrm{j}}\gamma\pi^e(a \mid o)b_V(a^+, \mathbf{O}^+)\nabla\Pr(\mathbf{O}^+, o \mid a, \mathbf{O}^-)\{b_W(A, \mathbf{O}^-)\odot\Pr(a, \mathbf{O}^-)\}$$
$$= (1-\gamma)^{-1}\sum_{a^+,o^+,o,a,o^-}\gamma\pi^e(a \mid o)b_V(a^+, o^+)\nabla\log P(o^+, o \mid a, o^-)b_W(a, o^-)\Pr(o^+, o, a, o^-).$$

Thus, $(1-\gamma)B_3$ is equal to

$$\gamma\mathbb{E}[b_W(A, O^-)\pi^e(A \mid O)\sum_{a^+}b_V(a^+, O^+)\nabla\log P(O^+, O \mid A, O^-)]$$
$$= \gamma\mathbb{E}[b_W(A, O^-)\pi^e(A \mid O)\sum_{a^+}b_V(a^+, O^+)\nabla\log P(R, O^+, O \mid A, O^-)]$$
$$= \gamma\mathbb{E}[b_W(A, O^-)\{\pi^e(A \mid O)\sum_{a^+}b_V(a^+, O^+) - \mathbb{E}[\pi^e(A \mid O)\sum_{a^+}b_V(a^+, O^+) \mid A, O^-]\}\nabla\log P(R, O^+, O \mid A, O^-)]$$
$$= \gamma\mathbb{E}[b_W(A, O^-)\{\pi^e(A \mid O)\sum_{a^+}b_V(a^+, O^+) - \mathbb{E}[\pi^e(A \mid O)\sum_{a^+}b_V(a^+, O^+) \mid A, O^-]\}\nabla\log P(R, O^+, O, A, O^-)].$$

Combining all things together, $(1-\gamma)(B_1+B_2+B_3)$ is

$$\mathbb{E}[b_W(A, O^-)\{R\pi^e(A \mid O) - \mathbb{E}[R\pi^e(A \mid O) \mid A, O^-]\}\nabla\log\Pr(O^+, R, O, A, O^-)]$$
$$+ \mathbb{E}[b_W(A, O^-)\{b_V(A, O) - \mathbb{E}[b_V(A, O) \mid A, O^-]\}\nabla\log\Pr(R, O^+, O, A, O^-)]$$
$$+ \gamma\mathbb{E}[b_W(A, O^-)\{\pi^e(A \mid O)\sum_{a^+}b_V(a^+, O^+) - \mathbb{E}[\pi^e(A \mid O)\sum_{a^+}b_V(a^+, O^+) \mid A, O^-]\}\nabla\log P(R, O^+, O, A, O^-)]$$
$$= \mathbb{E}[b_W(A, O^-)\{R\pi^e(A \mid O) + \gamma\pi^e(A \mid O)\sum_{a^+}b_V(a^+, O^+) - b_V(A, O)\}\nabla\log P(R, O^+, O, A, O^-)].$$

Hence, the following is ϕ_{EIF}:

$$\phi_{\mathrm{EIF}} = (1-\gamma)^{-1}b_W(A, O^-)\{R\pi^e(A \mid O) + \gamma\pi^e(A \mid O)\sum_{a^+}b_V(a^+, O^+) - b_V(A, O)\}.$$

□

Proof of Theorem 10. By some algebra, we have

$$J(f, g) - J(b_V, b_W) = B_1 + B_2 + B_3$$

where

$$B_1 = \{f(a, o^-) - b_W(a, o^-)\}\{\{\gamma\sum_{a'}\{g(a', O^+) - b_V(a', o^+)\}\}\pi^e(a \mid o) - g(a, o) + b_V(a', o^+)\},$$
$$B_2 = \mathbb{E}_{\tilde{o}\sim\nu_{\mathcal{O}}}[\sum_{a'}g(a, \tilde{o})] - \mathbb{E}_{\tilde{o}\sim\nu_{\mathcal{O}}}[\sum_{a'}b_V(a, \tilde{o})] + b_W(a, o^-)\{\gamma\sum_{a'}\{g(a', o^+) - b_V(a', o^+)\}\pi^e(a \mid o) - g(a, o) + b_V(a, o)\},$$
$$B_3 = \{b_W(a, o^-) - f(a, o^-)\}\{\{r + \gamma\sum_{a'}b_V(a', o^+)\}\pi^e(a \mid o) - b_V(a, o)\}\}.$$

Recall the estimator is constructed as

$$\hat{J}^*_{\mathrm{DR}} = 0.5\mathbb{E}_{\mathcal{D}_1}[J(\hat{b}_W^{(1)}, \hat{b}_V^{(1)})] + 0.5\mathbb{E}_{\mathcal{D}_2}[J(\hat{b}_W^{(2)}, \hat{b}_V^{(2)})].$$

We analyze $\mathbb{E}_{\mathcal{D}_1}[J(\hat{b}_W^{(1)}, \hat{b}_V^{(1)})] - J(\pi^e)$. This is expanded as

$$\begin{aligned}\sqrt{n}(\mathbb{E}_{\mathcal{D}_1}[J(\hat{b}_W^{(1)}, \hat{b}_V^{(1)})] - J(\pi^e)) &= \sqrt{n/n_1}\mathbb{G}_{\mathcal{D}_1}[J(\hat{b}_W^{(1)}, \hat{b}_V^{(1)}) - J(b_W, b_V)] \\ &+ \sqrt{n/n_1}\mathbb{G}_{\mathcal{D}_1}[J(b_W, b_V)] \\ &+ \sqrt{n/n_1}(\mathbb{E}[J(\hat{b}_W^{(1)}, \hat{b}_V^{(1)}) \mid \mathcal{D}_2] - J(\pi^e))\end{aligned}$$

where $\mathbb{G}_{\mathcal{D}_1} = \sqrt{n_1}(\mathbb{E}_{\mathcal{D}_1} - \mathbb{E})$. In the following, we analyze each term.

Analysis of $\sqrt{n/n_1}\mathbb{G}_{\mathcal{D}_1}[J(\hat{b}_W^{(1)}, \hat{b}_V^{(1)}) - J(b_W, b_V)]$. We show that the conditional variance given $\mathcal{D}_2$ is $o_p(1)$. Then, the rest of the argument is the same as the proof of (Kallus and Uehara, 2019, Theorem 7). This is proved by

$$\mathrm{var}[\sqrt{n_1}\mathbb{E}_{\mathcal{D}_1}[J(\hat{b}_W^{(1)}, \hat{b}_V^{(1)})] \mid \mathcal{D}_2] = \mathbb{E}[B_1^2 + B_2^2 + B_3^2 + 2B_1B_2 + 2B_1B_3 + 2B_1B_2 \mid \mathcal{D}_2] = o_p(1).$$

For example, $\mathbb{E}[B_3^2 \mid \mathcal{D}_2] = o_p(1)$ is proved by

$$\mathbb{E}[B_3^2 \mid \mathcal{D}_2] \lesssim \mathbb{E}[(\hat{b}_W - b_W)^2(A, O^-)]\mathbb{E}[\{\mathcal{T}\hat{b}_V\}^2(A, O^-)] = o_p(1).$$

Analysis of $\sqrt{n/n_1}(\mathbb{E}[J(\hat{b}_W^{(1)}, \hat{b}_V^{(1)}) \mid \mathcal{D}_2] - J(\pi^e))$. Here,we have

$$\begin{aligned}&\mathbb{E}[J(\hat{b}_W^{(1)}, \hat{b}_V^{(1)}) \mid \mathcal{D}_2] = |\mathbb{E}[B_1 + B_2 + B_3 \mid \mathcal{D}_2]| \\ &= |\mathbb{E}[B_3 \mid \mathcal{D}_2]| \leq \mathbb{E}[B_3^2 \mid \mathcal{D}_2]^{1/2} \leq \mathbb{E}[(\hat{b}_W - b_W)^2(A, O^-)]^{1/2}\mathbb{E}[\{\mathcal{T}\hat{b}_V\}^2(A, O^-)]^{1/2} \\ &= o_p(n_1^{-1/2}). \qquad \text{(From convergence rate condition)}\end{aligned}$$

Combine all things. Thus,

$$\sqrt{n}(\mathbb{E}_{\mathcal{D}_1}[J(\hat{b}_W^{(1)}, \hat{b}_V^{(1)})] - J(\pi^e)) = \sqrt{n/n_1}\mathbb{G}_{\mathcal{D}_1}[J(b_W, b_V)] + o_p(n^{-1/2}).$$

This immediately concludes

$$\sqrt{n}(\hat{J}^*_{DR} - J(\pi^e)) = \mathbb{G}_n[J(b_W, b_V)] + o_p(n^{-1/2})$$

where $\mathbb{G}_n = \sqrt{n}(\mathbb{E}_{\mathcal{D}} - \mathbb{E})$. □

E.5 Proof of Section 4

We denote $\mathbb{E}[b_V'^j(A_j, O_i) \mid A_j = a, S_j = s]$ by $Q^j(a, s)$.

Lemma 10.

$$Q^j(A_j, S_j) = \mathbb{E}[R_j\pi_j^e(A_j \mid O_j) \mid A_j, S_j] + \gamma\mathbb{E}\left[\sum_{a'} Q^{j+1}(a', S_{j+1})\pi_j^e(A_j \mid O_j) \mid A_j, S_j\right].$$

Proof. From the definition, we have

$$Q^j(A_j, S_j) = \mathbb{E}[R_j \pi_j^e(A_j \mid O_j) \mid S_j, A_j] + \mathbb{E}_{\pi^e}\left[\sum_{t=j+1}^{H-1} \gamma^{t-j} R_t \pi_j^e(A_j \mid O_j) \mid A_j, S_j\right].$$

Then,

$$\begin{aligned}
&\mathbb{E}_{\pi^e}\left[\sum_{t=j+1} \gamma^{t-j} R_t \pi_j^e(A_j \mid O_j) \mid A_j, S_j\right] = \mathbb{E}_{\pi^e}\left[\mathbb{E}_{\pi^e}\left[\sum_{t=j+1} \gamma^{t-j} R_t \mid A_j, S_{j+1}, O_{j+1}, O_j, S_j\right] \pi_j^e(A_j \mid O_j) \mid A_j, S_j\right] \\
&= \mathbb{E}\left[\sum_{a'} \mathbb{E}_{\pi^e}[\sum_{t=j+1} \gamma^{t-j} R_t \mid S_{j+1}, A_{j+1} = a', O_{j+1}, O_j] \pi_{j+1}^e(a' \mid O_{j+1}) \pi_j^e(A_j \mid O_j) \mid A_j, S_j\right] \\
&= \mathbb{E}\left[\sum_{a'} \mathbb{E}_{\pi^e}[\sum_{t=j+1} \gamma^{t-j} R_t \mid S_{j+1}, A_{j+1} = a'] \pi_{j+1}^e(a' \mid O_{j+1}) \pi_j^e(A_j \mid O_j) \mid A_j, S_j\right] \\
&= \mathbb{E}\left[\sum_{a'} \mathbb{E}_{\pi^e}[\sum_{t=j+1} \gamma^{t-j} R_t \mid S_{j+1}, A_{j+1} = a'] \mathbb{E}[\pi_{j+1}^e(a' \mid O_{j+1}) \mid S_{j+1}, A_{j+1} = a'] \pi_j^e(A_j \mid O_j) \mid A_j, S_j\right] \\
&= \mathbb{E}\left[\sum_{a'} \mathbb{E}_{\pi^e}[\sum_{t=j+1} \gamma^{t-j} R_t \pi_{j+1}^e(a' \mid O_{j+1}) \mid S_{j+1}, A_{j+1} = a'] \pi_j^e(A_j \mid O_j) \mid A_j, S_j\right] \\
&= \gamma \mathbb{E}\left[\sum_{a'} Q^{j+1}(a', S_{j+1}) \pi_j^e(A_j \mid O_j) \mid A_j, S_j\right].
\end{aligned}$$

□

Lemma 11. *[Learnable reward bridge functions are reward bridge functions]*

$$\mathbb{E}[b_V'^{[j]}(A_j, O_j) \mid A_j, \mathfrak{J}_{j-1}] = \mathbb{E}[R_j \pi_j^e(A_j \mid O_j) \mid A_j, \mathfrak{J}_{j-1}] + \gamma \mathbb{E}\left[\sum_{a'} b_V'^{[j+1]}(a', O_{j+1}) \pi_j^e(A_j \mid O_j) \mid A_j, \mathfrak{J}_{j-1}\right].$$

Proof. From Lemma 10, we have

$$\begin{aligned}
&\mathbb{E}[b_V'^{[j]}(A_j, O_j) \mid A_j, S_j] \\
&= \mathbb{E}[R_j \pi_j^e(A_j \mid O_j) \mid A_j, S_j] + \gamma \mathbb{E}\left[\sum_{a'} \mathbb{E}[b_V'^{[j+1]}(a', O_{j+1}) | S_{j+1}, O_j] \pi_j^e(A_j \mid O_j) \mid A_j, S_j\right] \\
&= \mathbb{E}[R_j \pi_j^e(A_j \mid O_j) \mid A_j, S_j] + \gamma \mathbb{E}\left[\sum_{a'} b_V'^{[j+1]}(a', O_{j+1}) \pi_j^e(A_j \mid O_j) \mid A_j, S_j\right].
\end{aligned}$$

Then, from $O_j, O_{j+1}, R_j \perp \mathfrak{J}_{j-1} \mid A_j, S_j$, the statement is concluded by

$$\begin{aligned}
\mathbb{E}[\mathbb{E}[b_V'^{[j]}(A_j, O_j) \mid A_j, S_j] \mid A_j, \mathfrak{J}_{j-1}] &= \mathbb{E}[b_V'^{[j]}(A_j, O_j) \mid A_j, \mathfrak{J}_{j-1}], \\
\mathbb{E}[\mathbb{E}[R_j \pi_j^e(A_j \mid O_j) \mid A_j, S_j] \mid A_j, \mathfrak{J}_{j-1}] &= \mathbb{E}[R_j \pi_j^e(A_j \mid O_j) \mid A_j, \mathfrak{J}_{j-1}], \\
\mathbb{E}\left[\mathbb{E}\left[\sum_{a'} b_V'^{[j+1]}(a', O_{j+1}) \pi_j^e(A_j \mid O_j) \mid A_j, S_j\right] \mid A_j, \mathfrak{J}_{j-1}\right] &= \mathbb{E}\left[\sum_{a'} b_V'^{[j+1]}(a', O_{j+1}) \pi_j^e(A_j \mid O_j) \mid A_j, \mathfrak{J}_{j-1}\right].
\end{aligned}$$

□

Lemma 12 (Learnable weight bridge functions are weight bridge functions)**.**

$$\mathbb{E}\left[b_W'^{[j]}(A_j, \mathfrak{J}_{j-1})\pi_j^e(A_j|O_j)f(O_{j+1}, a') - \mathrm{I}(A_{j+1} = a')b_W'^{[j+1]}(A_{j+1}, \mathfrak{J}_j)f(O_{j+1}, A_{j+1})\right] = 0,\ \forall f : \mathcal{O} \times \mathcal{A} \to \mathbb{R}.$$

and

$$\mathbb{E}\left[\sum_{a'} b_W'^{[j]}(A_j, \mathfrak{J}_{j-1})\pi_j^e(A_j|O_j)f(O_{j+1}, a') - b_W'^{[j+1]}(A_{j+1}, \mathfrak{J}_j)f(O_{j+1}, A_{j+1})\right] = 0,\ \forall f : \mathcal{O} \times \mathcal{A} \to \mathbb{R}.$$

Proof. We first prove the first statement:

$$\begin{aligned}
&\mathbb{E}\left[b_W'^{[j]}(A_j, \mathfrak{J}_{j-1})\pi_j^e(A_j|O_j)f(O_{j+1}, a') - \mathrm{I}(A_{j+1} = a')b_W'^{[j+1]}(A_{j+1}, \mathfrak{J}_j)f(O_{j+1}, A_{j+1})\right] \\
&= \mathbb{E}\left[\mathbb{E}[b_W'^{[j]}(A_j, \mathfrak{J}_{j-1}) \mid A_j, O_j, S_j, O_{j+1}]\pi_j^e(A_j|O_j)f(O_{j+1}, a')\right] \\
&- \mathbb{E}\left[\mathrm{I}(A_{j+1} = a')\mathbb{E}[b_W'^{[j+1]}(A_{j+1}, \mathfrak{J}_j) \mid O_{j+1}, A_{j+1}, S_{j+1}]f(O_{j+1}, A_{j+1})\right] \\
&= \mathbb{E}\left[\mu(S_j)\eta_j(A_j|O_j)f(O_{j+1}, a') - \mathrm{I}(A_{j+1} = a')\frac{\mu(S_{j+1})}{P_{\pi_b}(A_{j+1} \mid S_{j+1})}f(O_{j+1}, A_{j+1})\right] \\
&= \mathbb{E}_{\pi^e}[f(O_{j+1}, a')] - \mathbb{E}_{\pi^e}[f(O_{j+1}, a')] \\
&= 0.
\end{aligned}$$

Then, we can prove the second statement by taking summation over the action space.

□

Lemma 13 (Identification formula with unlearnable bridge functions)**.** *Assume the existence of bridge functions. Then, let $b_V' = \{b_V'^{[j]}\}, b_W' = \{b_W'^{[j]}\}$ be any bridge functions.*

$$J(\pi^e) = \mathbb{E}\left[\sum_j \gamma^j b_W'^{[j]}(A_j, \mathfrak{J}_{j-1})\pi_j^e(A_j \mid O_j)R_j\right], \quad J(\pi^e) = \mathbb{E}\left[\sum_a b_V'^{[0]}(a, O_0)\right].$$

Proof. We prove the first formula:

$$\begin{aligned}
\mathbb{E}\left[\sum_j \gamma^j b_W'^{[j]}(A_j, \mathfrak{J}_{j-1})\pi_j^e(A_j \mid O_j)R_j\right] &= \mathbb{E}\left[\sum_j \gamma^j \mathbb{E}[b_W'^{[j]}(A_j, \mathfrak{J}_{j-1}) \mid A_j, S_j]\pi_j^e(A_j \mid O_j)R_j\right] \\
&= \mathbb{E}\left[\sum_j \gamma^j \frac{\mu_j(S_j)}{P_{\pi_b}(A_j \mid S_j)}\pi_j^e(A_j \mid O_j)R_j\right] \\
&\qquad \text{(Definition of bridge functions)} \\
&= J(\pi^e).
\end{aligned}$$

We prove the second formula:

$$\begin{aligned}
&\mathbb{E}\left[\sum_a b_V'^{[0]}(a, O_0)\right] = \mathbb{E}\left[\mathbb{E}[\sum_a b_V'^{[0]}(a, O_0) \mid S_0]\right] = \mathbb{E}\left[\mathbb{E}[\sum_a b_V'^{[0]}(a, O_0) \mid S_0, A_0 = a]\right] \\
&= \mathbb{E}\left[\sum_a \mathbb{E}_{\pi^e}[\sum_{t=0}^{H-1} \gamma^t R_t \pi_0^e(a \mid O_0) \mid S_0, A_0 = a]\right] && \text{(Definition of bridge functions)} \\
&= \mathbb{E}\left[\sum_a \mathbb{E}_{\pi^e}[\sum_{t=0}^{H-1} \gamma^t R_t \mid S_0, A_0 = a]\mathbb{E}[\pi_0^e(a \mid O_0) \mid S_0, A_0 = a]\right] \\
&= \mathbb{E}\left[\sum_a \mathbb{E}_{\pi^e}[\sum_{t=0}^{H-1} \gamma^t R_t \mid S_0, A_0 = a]\pi_0^e(a \mid O_0)\right] \\
&= J(\pi^e).
\end{aligned}$$

□

Lemma 14 (Final identification formula)**.**

- *Assume the existence of bridge functions. Let $b_V = \{b_V^{[j]}\}, b_W = \{b_W^{[j]}\}$ be any learnable bridge functions.*

$$J(\pi^e) = \mathbb{E}[\sum_a b_V^{[0]}(a, O_0)], \quad J(\pi^e) = \mathbb{E}\left[\sum_{j=0}^{H-1} \gamma^j b_W^{[j]}(A_j, \mathfrak{J}_{j-1}) \pi_j^e(A_j \mid O_j) R_j\right].$$

- *Assume the existence of learnable value bridge functions and the completeness:*

$$\mathbb{E}[g(S_j, A_j) \mid A_j, \mathfrak{J}_{j-1}] = 0 \implies g(\cdot) = 0.$$

Then,

$$J(\pi^e) = \mathbb{E}[\sum_a b_V^{[0]}(a, O_0)].$$

Proof. We provide the proof without completeness. The proof with completeness is given in Theorem 14. We prove

the first formula:

$$
\begin{aligned}
J(\pi^e) &= \mathbb{E}\left[\sum_j \gamma^j b_W'^{[j]}(A_j, \mathfrak{J}_{j-1})\pi_j^e(A_j \mid O_j)R_j\right] && \text{(From Lemma 13)} \\
&= \mathbb{E}\left[\sum_j \gamma^j b_W'^{[j]}(A_j, \mathfrak{J}_{j-1})\mathbb{E}[\pi_j^e(A_j \mid O_j)R_j \mid A_j, \mathfrak{J}_{j-1}]\right] \\
&= \mathbb{E}\left[\sum_j \gamma^j b_W'^{[j]}(A_j, \mathfrak{J}_{j-1})\{\gamma \sum_a \mathbb{E}[b_V^{[j+1]}(a, O_{j+1})\pi_j^e(A_j \mid O_j) \mid A_j, \mathfrak{J}_{j-1}] - \mathbb{E}[b_V^{[j]}(A_j, O_j) \mid A_j, \mathfrak{J}_{j-1}]\}\right] \\
& && \text{(b_V are value bridge functions.)} \\
&= \sum_{j=1}^{H-1} \gamma^j \mathbb{E}\left[b_W'^{j-1}(A_{j-1}, \mathfrak{J}_{j-2})\pi_{j-1}^e(A_{j-1} \mid O_{j-1}) \sum_a b_V^{[j]}(a, O_j) - b_W'^{[j]}(A_j, \mathfrak{J}_{j-1})b_V^{[j]}(A_j, O_j)\right] \\
&+ \mathbb{E}[\sum_a b_V^{[0]}(a, O_0)] \\
&= \mathbb{E}[\sum_a b_V^{[0]}(a, O_0)]. && \text{(Value bridge functions are learnable bridge functions.)}
\end{aligned}
$$

Next, we prove the second formula:

$$
\begin{aligned}
J(\pi^e) &= \mathbb{E}[\sum_a b_V'^{[0]}(a, O_0)] && \text{(From Lemma 13)} \\
&= \sum_{j=1}^{H-1} \gamma^j \mathbb{E}\left[b_W^{[j-1]}(A_{j-1}, \mathfrak{J}_{j-2})\pi_{j-1}^e(A_{j-1} \mid O_{j-1}) \sum_a b_V'^{[j]}(a, O_j) - b_W^{[j]}(A_j, \mathfrak{J}_{j-1})b_V'^{[j]}(A_j, O_j)\right] \\
& && \text{(Definition of weight bridge functions)} \\
&+ \mathbb{E}[\sum_a b_V'^{[0]}(a, O_0)] \\
&= \mathbb{E}\left[\sum_j \gamma^j b_W^{[j]}(A_j, \mathfrak{J}_{j-1})\{\gamma \sum_a \mathbb{E}[b_V'^{(j+1)}(a, O_{j+1})\pi_j^e(A_j \mid O_j) \mid A_j, \mathfrak{J}_{j-1}] - \mathbb{E}[b_V'^{j}(A_j, O_j) \mid A_j, \mathfrak{J}_{j-1}]\}\right] \\
&= \mathbb{E}\left[\sum_j \gamma^j b_W^{[j]}(A_j, \mathfrak{J}_{j-1})\mathbb{E}[\pi_j^e(A_j \mid O_j)R_j \mid A_j, \mathfrak{J}_{j-1}]\right] \\
& && \text{(Reward bridge functions are learnable reward bridge functions)} \\
&= \mathbb{E}\left[\sum_j \gamma^j b_W^{[j]}(A_j, \mathfrak{J}_{j-1})\pi_j^e(A_j \mid O_j)R_j\right].
\end{aligned}
$$

□

Proof of Lemma 4. This is proved in Nair and Jiang (2021, Theorem 3). Especially, our formula is their specific formula when using left singular vectors for M. We refer the readers to read their proof. □

Next, we prove the doubly robust property.

Proof of Theorem 13. The first statement $J(\pi^e) = \mathbb{E}[J(b_V, g)]$ is proved by the definition of reward bridge functions:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_a b_V^{[0]}(a, O_0) + \sum_{k=0}^{H-1} \gamma^k g^{[k]}(A_k, \mathfrak{J}_{k-1}) \left(\pi_k^e(A_k \mid O_k)\{R_k + \gamma \sum_{a^+} b_V^{[k+1]}(a^+, O_{k+1})\} - b_V^{[k]}(A_k, O_k)\right)\right] \\
&= \mathbb{E}\left[\sum_a b_V^{[0]}(a, O_0) + \sum_{k=0}^{H-1} \gamma^k \hat{g}^{[k]}(A_k, \mathfrak{J}_{k-1}) \mathbb{E}\left(\pi_k^e(A_k \mid O_k)\{R_k + \gamma \sum_{a^+} \hat{b}_V^{[k+1]}(a^+, O_{k+1})\} - b_V^{[k]}(A_k, O_k) \mid A_k, \mathfrak{J}_{k-1}\right)\right] \\
&= \mathbb{E}\left[\sum_a b_V^{[0]}(a, O_0)\right] && \text{(From the definition of reward bridge functions)} \\
&= J(\pi^e). && \text{(From identification results, Theorem 14)}
\end{aligned}
$$

The second statement $J(\pi^e) = \mathbb{E}[J(f, b_W)]$ is proved by the definition of weight bridge functions:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_a f^{[0]}(a, O_0) + \sum_{k=0}^{H-1} \gamma^k b_W^{[k]}(A_k, \mathfrak{J}_{k-1}) \left(\pi_k^e(A_k \mid O_k)\{R_k + \gamma \sum_{a^+} f^{[k+1]}(a^+, O_{k+1})\} - f^{[k]}(A_k, O_k)\right)\right] \\
&= \mathbb{E}\left[\sum_{k=0}^{H-1} \gamma^k b_W^{[k]}(A_k, \mathfrak{J}_{k-1}) \pi_k^e(A_k \mid O_k) R_k\right] + \\
&+ \mathbb{E}\left(\sum_{k=1}^{H-1} \gamma^k b_W^{[k-1]}(A_{k-1}, \mathfrak{J}_{k-2}) \pi_{k-1}^e(A_{k-1} \mid O_{k-1}) \sum_{a^+} f^{[k]}(a^+, O_k) - \gamma^k b_W^{[k]}(A_k, \mathfrak{J}_{k-1}) f^{[k]}(A_k, O_k)\right) \\
&= \mathbb{E}\left[\sum_{k=0}^{H-1} \gamma^k b_W^{[k]}(A_k, \mathfrak{J}_{k-1}) \pi_k^e(A_k \mid O_k) R_k\right] && \text{(From the definition of weight bridge functions)} \\
&= J(\pi^e). && \text{(From identification results, Theorem 14)}
\end{aligned}
$$

□

Prof of Theorem 14.

With Completeness From the definition of bridge functions:

$$
\mathbb{E}\left[\pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) R_j + \sum_a \gamma b_V^{[j+1]}(a, O_{j+1}, \tilde{\mathfrak{J}}_j) \pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) - b_V^{[j]}(A_j, O_j, \tilde{\mathfrak{J}}_{j-1}) \mid A_j, \mathfrak{J}_{j-1}\right] = 0.
$$

Using the completeness, we have

$$
\mathbb{E}\left[\pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) R_j + \sum_a \gamma b_V^{[j+1]}(a, O_{j+1}, \tilde{\mathfrak{J}}_j) \pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) - b_V^{[j]}(A_j, O_j, \tilde{\mathfrak{J}}_{j-1}) \mid S_j, A_j, \tilde{\mathfrak{J}}_{j-1}\right] = 0.
$$

Now, we are ready to prove the main statement. We denote $\mu_j(S_j, \tilde{\mathfrak{J}}_{j-1}) = \mathbb{E}[\prod_{k=0}^{j-1} \frac{\pi_k^e(A_k|O_k, \mathfrak{J}_{k-1})}{\pi_b(A_k|S_k)} \mid S_j, \tilde{\mathfrak{J}}_{j-1}]$.

Then, we have

$$
\begin{aligned}
J(\pi^e) &= \mathbb{E}\left[\sum_j \gamma^j \frac{\mu_j(S_j, \tilde{\mathfrak{J}}_{j-1})}{\pi_b(A_j \mid S_j)} \pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) R_j\right] \\
&= \mathbb{E}\left[\sum_j \gamma^j \frac{\mu_j(S_j, \tilde{\mathfrak{J}}_{j-1})}{\pi_b(A_j \mid S_j)} \mathbb{E}\left[\pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) R_j \mid S_i, A_j, \tilde{\mathfrak{J}}_{j-1}\right]\right] \\
&= \mathbb{E}\left[\sum_j \gamma^j \frac{\mu_j(S_j, \tilde{\mathfrak{J}}_{j-1})}{\pi_b(A_j \mid S_j)} \{\gamma \sum_a \mathbb{E}[b_V^{[j+1]}(a, O_{j+1}, \tilde{\mathfrak{J}}_j) \pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_j) \mid S_j, A_j, \tilde{\mathfrak{J}}_{j-1}] - \mathbb{E}[b_V^{[j]}(A_j, O_j, \tilde{\mathfrak{J}}_{j-1}) \mid S_j, A_j, \tilde{\mathfrak{J}}_{j-1}]\}\right] \\
&= \sum_{j=1}^{H-1} \gamma^j \mathbb{E}\left[\frac{\mu_{j-1}(S_{j-1}, \tilde{\mathfrak{J}}_{j-2})}{\pi_b(A_{j-1} \mid S_{j-1})} \pi_{j-1}^e(A_{j-1} \mid O_{j-2}, \tilde{\mathfrak{J}}_{j-2}) \sum_a b_V^{[j]}(a, O_j, \tilde{\mathfrak{J}}_{j-1}) - \frac{\mu_j(S_j, \tilde{\mathfrak{J}}_{j-1})}{\pi_b(A_j \mid S_j)} b_V^{[j]}(A_j, O_j, \tilde{\mathfrak{J}}_{j-1})\right] \\
&+ \mathbb{E}[\sum_a b_V^{[0]}(a, O_0)] = \mathbb{E}[\sum_a b_V^{[0]}(a, O_0)].
\end{aligned}
$$

Without Completeness From the definition of bridge functions:

$$
\mathbb{E}\left[\pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) R_j + \sum_a \gamma b_V^{[j+1]}(a, O_{j+1}, \tilde{\mathfrak{J}}_j) \pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) - b_V^{[j]}(A_j, O_j, \tilde{\mathfrak{J}}_{j-1}) \mid A_j, \tilde{\mathfrak{J}}_{j-1}\right] = 0.
$$

Now, we are ready to prove the main statement. We denote $\mu_{j-1}(S_j, \tilde{\mathfrak{J}}_{j-1}) = \mathbb{E}[\prod_{k=0}^{j-1} \frac{\pi_k^e(A_k|O_k, \tilde{\mathfrak{J}}_{k-1})}{\pi_b(A_k|S_k)} \mid S_j, A_j, \mathfrak{J}_{j-1}]$. Then, we have

$$
\begin{aligned}
J(\pi^e) &= \mathbb{E}\left[\sum_j \gamma^j \frac{\mu_j(S_j, \tilde{\mathfrak{J}}_{j-1})}{\pi_b(A_j \mid S_j)} \pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) R_j\right] \\
&= \mathbb{E}\left[\sum_j \gamma^j \left\{\mathbb{E}[b_W^{[j]}(A_j, \mathfrak{J}_{j-1}) \mid S_j, A_j, \tilde{\mathfrak{J}}_{j-1}]\right\} \pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) R_j\right] \\
&= \mathbb{E}\left[\sum_j \gamma^j b_W^{[j]}(A_j, \mathfrak{J}_{j-1}) \pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) R_j\right] \\
&= \mathbb{E}\left[\sum_j \gamma^j b_W^{[j]}(A_j, \mathfrak{J}_{j-1}) \{\gamma \sum_a \mathbb{E}[b_V^{[j+1]}(a, O_{j+1}, \tilde{\mathfrak{J}}_j) \pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) - b_V^{[j]}(A_j, O_j, \tilde{\mathfrak{J}}_{j-1}) \mid A_j, \mathfrak{J}_{j-1}]\}\right] \\
&= \mathbb{E}\left[\sum_j \gamma^j \frac{\mu_j(S_j, \tilde{\mathfrak{J}}_{j-1})}{\pi_b(A_j \mid S_j)} \{\gamma \sum_a \mathbb{E}[b_V^{[j+1]}(a, O_{j+1}, \tilde{\mathfrak{J}}_j) \pi_j^e(A_j \mid O_j, \tilde{\mathfrak{J}}_{j-1}) - b_V^{[j]}(A_j, O_j, \tilde{\mathfrak{J}}_{j-1}) \mid A_j, \mathfrak{J}_{j-1}]\}\right] \\
&= \sum_{j=1}^{H-1} \gamma^j \mathbb{E}\left[\frac{\mu_{j-1}(S_{j-1}, \tilde{\mathfrak{J}}_{j-2})}{\pi_b(A_{j-1} \mid S_{j-1})} \pi_{j-1}^e(A_{j-1} \mid O_{j-1}, \tilde{\mathfrak{J}}_{j-1}) \sum_a b_V^{[j]}(a, O_j, \tilde{\mathfrak{J}}_{j-1}) - \frac{\mu_j(S_j, \tilde{\mathfrak{J}}_{j-1})}{\pi_b(A_j \mid S_j)} b_V^{[j]}(A_j, O_j, \tilde{\mathfrak{J}}_{j-1})\right] \\
&+ \mathbb{E}[\sum_a b_V^{[0]}(a, O_0)] = \mathbb{E}[\sum_a b_V^{[0]}(a, O_0)].
\end{aligned}
$$

□

E.6 Proof of Appendix C

Proof of Theorem 16. From Hoeffding inequality, we use with probability $1-\delta$:

$$\forall f \in \mathcal{V}, \forall g \in \mathcal{V}^\dagger; |(\mathbb{E} - \mathbb{E}_\mathcal{D}) \left[f(A, O^-) \mathcal{Z} g(A, O^-) \right]| \leq c\{\max(1, C_\mathcal{V}, C_{\mathcal{V}^\dagger})\} \sqrt{\log(|\mathcal{V}||\mathcal{V}^\dagger|/\delta)/n}.$$

Recall $\mathcal{Z} f = R\pi^e(a \mid o) + \sum_{a'} f(a', O^+)\pi^e(a \mid o)$. Hereafter, we condition this event.

In this proof, we often use the following:

$$\mathbb{E}[f(A, O^-) \mathcal{Z} g(A, O^-)] = \mathbb{E}[f(A, O^-) \mathcal{T} g(A, O^-)].$$

Recall $\mathcal{T} g = \mathbb{E}[R\pi^e(A \mid O) + \sum_{a'} f(a', O^+)\pi^e(a \mid O) \mid A = a, O^- = o]$.

As a first step, we have

$$|J(\pi^e) - \hat{J}_{\mathrm{VM}}| \leq \left| (1-\gamma)^{-1} \mathbb{E} \left[b_W(A, O^-) \mathcal{T} \hat{b}_V(A, O^-) \right] \right|.$$

Then,

$$\begin{aligned}
\left| \mathbb{E} \left[b_W(A, O^-) \mathcal{T} \hat{b}_V(A, O^-) \right] \right| &\leq \left| (\mathbb{E} - \mathbb{E}_\mathcal{D}) \left[b_W(A, O^-) \mathcal{Z} \hat{b}_V(A, O^-) \right] \right| + \left| \mathbb{E}_\mathcal{D} \left[b_W(A, O^-) \mathcal{Z} \hat{b}_V(A, O^-) \right] \right| \\
&\leq c\sqrt{\log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)/n} + \sup_{f \in \mathcal{V}^\dagger} \left| \mathbb{E}_\mathcal{D} \left[f(A, O^-) \mathcal{Z} \hat{b}_V(A, O^-) \right] \right| \qquad (b_W \in \mathcal{V}^\dagger) \\
&\leq c\sqrt{\log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)/n} + \sup_{f \in \mathcal{V}^\dagger} \left| \mathbb{E}_\mathcal{D} \left[f(A, O^-) \mathcal{Z} b_V(A, O^-) \right] \right| \qquad (b_V \in \mathcal{V}) \\
&\leq 2c\sqrt{\log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)/n}.
\end{aligned}$$

In the final line, we use

$$\mathbb{E} \left[f(A, O^-) \mathcal{T} b_V(A, O^-) \right] = 0, \quad \forall f : \mathcal{A} \times \mathcal{O} \to \mathbb{R};$$

thus,

$$\begin{aligned}
&\sup_{f \in \mathcal{V}^\dagger} \left| \mathbb{E}_\mathcal{D} \left[f(A, O^-) \mathcal{Z} b_V(A, O^-) \right] \right| \\
&= \left| \mathbb{E}_\mathcal{D} \left[f^\diamond(A, O^-) \mathcal{Z} b_V(A, O^-) \right] \right| \qquad (f^\diamond = \max_{f \in \mathcal{V}^\dagger} |\mathbb{E}_\mathcal{D} \left[f(A, O^-) \mathcal{Z} b_V(A, O^-) \right]|) \\
&\leq \left| \mathbb{E}[\left[f^\diamond(A, O^-) \mathcal{Z} b_V(A, O^-) \right]| + c\sqrt{\log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)/n} \right. \\
&= \left| \mathbb{E}[\left[f^\diamond(A, O^-) \mathcal{T} b_V(A, O^-) \right]| + c\sqrt{\log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)/n} = c\sqrt{\log(|\mathcal{V}||\mathcal{V}^\dagger|c/\delta)/n}. \right.
\end{aligned}$$

□