A Additional Literature review

Minimax learning. Minimax learning (also known as adversarial learning) has been widely applied to a large variety of problems ranging from instrumental variable estimation [BKS19, DLMS20] to policy learning/evaluation in contextual bandits/MDPs [HW17, CNS18, FS19, FLL19, LLTZ18, CNSS20, UIJ+21]. For example, in OPE problems with fully-observable environments, minimax learning methods have been developed to learn q-functions and marginal ratios that are characterized as solutions to certain conditional moment equations [UHJ20]. The solutions to these conditional moment equations are uniquely defined. On the other hand, our case is more challenging since the solutions to conditional moment equations are not uniquely defined. Although the solutions are *not* uniquely defined and hence cannot be identified, our estimands, i.e., policy values, *can* be still identified. This requires significantly distinctive analysis, which is not seen in standard IV settings or MDP settings.

B Comparison to Analogue of Future-dependent Value Functions

Analogs of our future-dependent value functions have been introduced in confounded contextual bandits and confounded POMDPs as bridge functions [MGT18, CPS+20, TSM20, SUHJ22, BK21]. These works consider confounded settings where actions depend on unobserved states and introduce bridge functions to deal with confounding. Instead, we introduce future-dependent value functions to deal with the curse of the horizon while there is no confounding issue. Existing definitions of bridge functions in confounded POMDPs do not work in standard POMDPs. In the definition of existing bridge functions, behavior policies cannot depend on observations O since observations O are used as so-called negative controls, which do not affect action A and are not affected by action A. In our setting, O does *not* serve as negative controls unlike their works since A clearly depends on O. Instead, O just play a role in covariates. See Figure 3. Due to this fact, we can further add F as input domains of future-dependent value functions, unlike bridge functions by regarding F as just covariates. This is impossible in the definition of existing bridge functions without further assumptions as mentioned in [NJ21]. In this sense, our setting does *not* fall into the category of the so-called proximal causal inference framework. At the same time, our definition does not work in these confounded settings since Definition 2 explicitly includes behavior policies.

We finally remark the observation that history can serve as an instrumental variable in POMDPs is mentioned in [HDG15, VSH+16]. However, they did not propose future-dependent value functions; their goal is to learn system dynamics.

C Off-Policy Evaluation for Memory-Based Policies

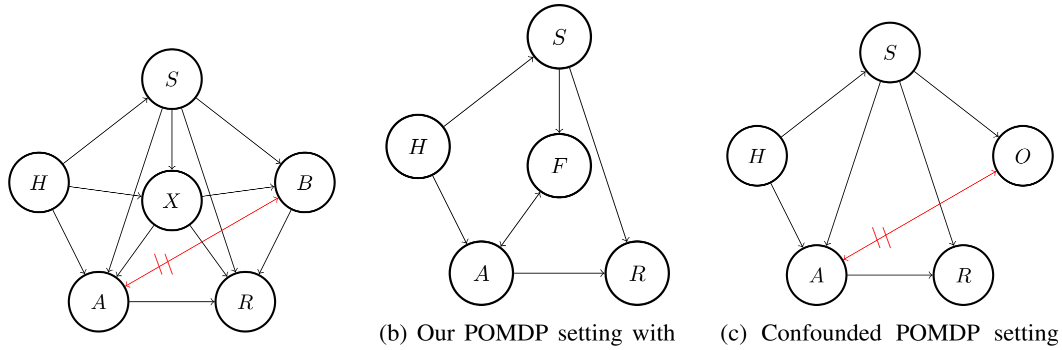
So far, we have discussed how to evaluate memoryless policies to simplify the notation. In this section, we will now turn our attention to the evaluation of memory-based policies.

C.1 Settings

We consider M-memory policies $\pi : \mathcal{Z} \times \mathcal{O} \to \Delta(\mathcal{A})$ that are functions of the current observation O_t and past observation-action pairs at time point $t, t-1, \cdots, t_M$ denoted by $Z_t = (O_{t-M:t-1}, A_{t-M:t-1}) \in \mathcal{Z} = \mathcal{O}^M \times \mathcal{A}^M$, for some integer $M > 1$. We assume the existence of M observation pairs obtained prior to the initial time point (denoted by Z_0). Following an M-memory policy π, the data generating process can be described as follows. First, Z_0 and S_0 are generated according to some initial distribution $\nu_{\bar{\mathcal{S}}} \in \Delta(\bar{\mathcal{S}})$ where $\bar{\mathcal{S}} = \mathcal{Z} \times \mathcal{S}$. Next, the agent observes $O_0 \sim \mathbb{O}(\cdot \mid S_0)$, executes the initial action $A_0 \sim \pi(\cdot \mid Z_0, O_0)$, receives a reward $r(S_0, A_0)$, the environment transits to the next state $S_1 \sim \mathbb{T}(\cdot \mid S_0, A_0)$, and this process repeats. See Figure 4 for a graphical illustration of the data-generating process. We assume that both the behavior and evaluation policies are M-memory.

Our goal is to estimate a policy value $J(\pi^e)$ for an M-meory evaluation policy π^e. Toward this end, we define a state-value function under π^e:

$$V^{\pi^e}(z, s) := \mathbb{E}_{\pi^e}[\sum_{k=0}^{\infty} \gamma^k R_k \mid Z_0 = z, S_0 = s]$$



(a) Standard negative control setting

(b) Our POMDP setting with $\gamma = 0$

(c) Confounded POMDP setting with $\gamma = 0$

Figure 3: Comparisons of three DAGs. $\backslash\backslash$ means no arrows. For simplicity, we consider the case with $\gamma = 0$, i.e., we do not need the next transitions. The first graph is a graph used in [CPS+20]. Note H, B, X are an action negative control, a reward negative control, and a covariate, respectively. We need $(H, A) \perp B \mid S, X$ and $H \perp Y \mid S, X, A$. The graph is one instance satisfying this condition. The second graph corresponds to the contextual bandit version ($\gamma = 0$) of our setting. Future proxies F just serve as a covariate and H serves as an action negative control. There are no nodes that correspond to reward negative controls. The third graph corresponds to the contextual bandit version ($\gamma = 0$) of confounded POMDPs [TSM20, SUHJ22, BK21]. A node O corresponds to a reward negative control, and H corresponds to an action negative control that satisfies $(H, A) \perp O \mid S$. Thus, O cannot include futures proxies (F) since then we cannot ensure $(H, A) \perp F \mid S$ since there is an arrow from A to F.

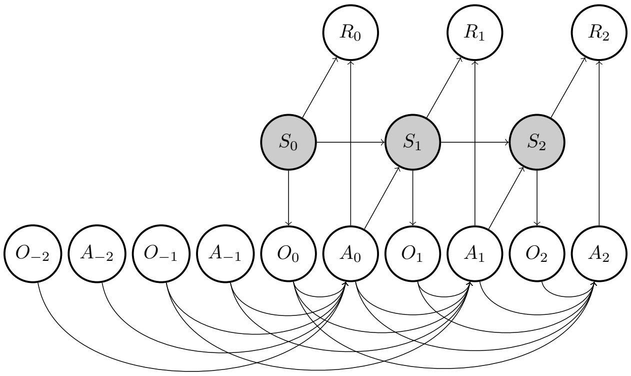


Figure 4: POMDPs when $M = 2$. Note S_0, O_0, A_0, R_0 correspond to (S, O, A, R), respectively, and $(O_{-2}, A_{-2}, O_{-1}, A_{-1}) = Z_0$. We cannot observe S in the offline data.

for any $z \in \mathcal{Z}, s \in \mathcal{S}$. Compared to the memory-less case, the input additionally includes z. Let $\bar{S}_t = (Z_t, S_t)$, and $d_t^{\pi^e}(\cdot)$ be the marginal distribution of $\bar{S}_t$ under the policy π^e.

Next, we explain how the offline data is collected when behavior policies are M-memory. Specifically, the dataset $\mathcal{D}_{\rm tra}$ consists of n data tuples $\{(H^{(i)}, O^{(i)}, A^{(i)}, R^{(i)}, F'^{(i)})\}_{i=1}^{N}$. We use (H, O, A, R, F') to denote a generic history-observation-action-reward-future tuple where H denotes the M_H-step historical observations obtained prior to the observation O and F' denotes the M_F-step future observations after (O, A) for some integers $M_H > M$ and $M_F \geq 1$. Hence, given some time step t in the trajectory data, we set $(O, A, R) = (O_t, A_t, R_t)$,

$$H = (O_{-M_H:t-1}, A_{-M_H:t-1}) \text{ and } F' = (O_{t+1:t+M_F}, A_{t+1:t+M_F-1}).$$

We additionally set $F = (O_{t:t+M_F-1}, A_{t:t+M_F-2})$. We use the prime symbol ' to represent the next time step. Then, $Z' = (O_{t-M+1:t}.A_{t-M+1:t})$. See Figure 5 for details when we set $t = 0$.

Throughout this paper, uppercase letters such as (H, S, O, A, R, S', F') are reserved for *random variables* and lower case letters such as (h, s, o, a, r, s', f') are reserved for their *realizations*. For

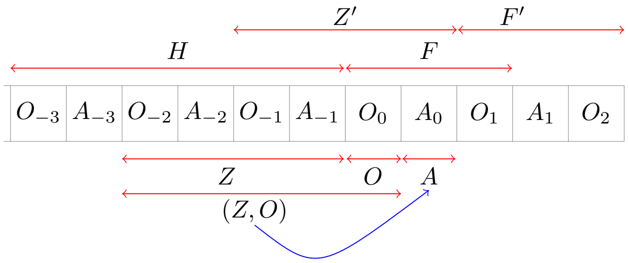


Figure 5: Case with $M_H = 3, M = 2, M_F = 2$. A 2-memory policy determines action A based on (Z, O).

Algorithm 2 Minimax OPE on POMDPs

```
Require: Dataset $\mathcal{D}$, function classes $\mathcal{Q} \subset [\bar{\mathcal{F}} \to \mathbb{R}], \Xi \subset [\mathcal{H} \to \mathbb{R}]$, hyperparameter $\lambda \geq 0$
1: $\hat{b}_V = \mathrm{argmin}_{q \in \mathcal{Q}} \max_{\xi \in \Xi} \mathbb{E}_{\mathcal{D}}[\{\mu(Z, A, O)\{R + \gamma q(\bar{F}')\} - q(\bar{F})\}\xi(H) - \lambda \xi^2(H)]$.
2: return $\hat{J}_{\mathrm{VM}} = \mathbb{E}_{\mathcal{D}_{\mathrm{ini}}}[\hat{b}_V(\bar{f})]$
```

simplicity, we impose the stationarity assumption, i.e., the marginal distributions of (H, S, F) and (H', S', F') are the same.

The dataset $\mathcal{D}_{\mathrm{ini}}$ consists of N' data tuples $\{Z_0^{(i)}, O_{0:M_F-1}^{(i)}, A_{0:M_F-1}^{(i)}\}_{i=1}^{N'}$ which is generated as follows: $\bar{S}_0 \sim \nu_{\bar{\mathcal{S}}}, O_0 \sim \mathbb{O}(\cdot \mid S_0), A_0 \sim \pi^b(\cdot \mid Z_0, O_0), S_1 \sim \mathbb{T}(\cdot \mid S_0, A_0), \cdots$, until we observe $O_{M_F-1}^{(i)}$ and $A_{M_F-1}^{(i)}$. We denote its distribution over $\bar{\mathcal{F}} = \mathcal{Z} \times \mathcal{F}$ by $\nu_{\bar{\mathcal{F}}}(\cdot)$.

Notation. We denote the domain of Z by $\mathcal{Z} = \mathcal{O}^M \times \mathcal{A}^M$. We define $\bar{S} = (Z, S), \bar{F} = (Z, F)$.

C.2 Required changes in Section 3

Every definition and statement holds by replacing $F, S, F', \mathcal{F}, \mathcal{S}$ with $\bar{F}, \bar{S}, \bar{F}', \bar{\mathcal{F}}, \bar{\mathcal{S}}$, respectively. For completeness, we show these definitions and theorems tailored to M-memory policies.

Definition 4 (Future-dependent value functions)**.** *Future-dependent value functions* $g_V \in [\bar{\mathcal{F}} \to \mathbb{R}]$ *are defined such that the following holds almost surely,*

$$\mathbb{E}[g_V(\bar{F}) \mid \bar{S}] = V^{\pi^e}(\bar{S}).$$

Definition 5 (Learnable future-dependent value functions)**.** *Define* $\mu(Z, O, A) := \pi^e(A \mid Z, O)/\pi^b(A \mid Z, O)$. *Learnable future-dependent value functions* $b_V \in [\bar{\mathcal{F}} \to \mathbb{R}]$ *are defined such that the following holds almost surely,*

$$0 = \mathbb{E}\left[\mu(Z, O, A)\{R + \gamma b_V(\bar{F}')\} - b_V(\bar{F}) \mid H\right]. \tag{10}$$

We denote the set of solutions by $\mathcal{B}_V$.

Theorem 4 (Identification Theorem)**.** *Suppose (a) the existence of learnable future-dependent value functions (need not be unique); (b) the invertiblity condition, i.e., any* $g : \bar{\mathcal{S}} \to \mathbb{R}$ *that satisfies* $\mathbb{E}[g(\bar{S}) \mid H] = 0$ *must also satisfy* $g(\bar{S}) = 0$ *(i.e.,* $g(\bar{s}) = 0$ *for almost every* $\bar{s}$ *that belongs to the support of* $\bar{S}$*), (c) the overlap condition* $w_{\pi^e}(\bar{S}) < \infty, \mu(Z, O, A) < \infty$*. Then, for any* $b_V \in \mathcal{B}_V$*, we have*

$$J(\pi^e) = \mathbb{E}_{\bar{f} \sim \nu_{\bar{\mathcal{F}}}}[b_V(\bar{f})]. \tag{11}$$

C.3 Required changes in Section 4

Every algorithm holds by replacing $F, S, F', \mathcal{F}, \mathcal{S}$ with $\bar{F}, \bar{S}, \bar{F}', \bar{\mathcal{F}}, \bar{\mathcal{S}}$, respectively. For completeness, we show the modified version of Algorithm 1 in Algorithm 2.

C.4 Required change in Section 5

We present the modified version of Theorem 3 tailored to M-memory policies.

Definition 6 (Bellman operators)**.** *The Bellman residual operator onto the history is defined as*

$$\mathcal{T} : [\bar{\mathcal{F}} \to \mathbb{R}] \ni q(\cdot) \mapsto \mathbb{E}[\mu(Z,O,A)\{R + \gamma q(\bar{F}')\} - q(\bar{F}) \mid H = \cdot],$$

and the Bellman residual error onto the history is defined as $\mathbb{E}[(\mathcal{T}q)^2(H)]$*. Similarly, the Bellman residual operator onto the latent state,* $\mathcal{T}^S$ *is defined as*

$$\mathcal{T}^S : [\bar{\mathcal{F}} \to \mathbb{R}] \ni q(\cdot) \mapsto \mathbb{E}[\mu(Z,O,A)\{R + \gamma q(\bar{F}')\} - q(\bar{F}) \mid \bar{S} = \cdot],$$

and the Bellman residual error onto the latent state is defined as $\mathbb{E}[\{\mathcal{T}^S(q)\}^2(\bar{S})]$*.*

Theorem 5 (Finite sample property of $\hat{b}_V$)**.** *Set* $\lambda > 0$*. Suppose (5a)* $\mathcal{B}_V \cap \mathcal{Q} \neq 0$ *(realizability) and (5b)* $\mathcal{T}\mathcal{Q} \subset \Xi$ *(Bellman completeness). With probability* $1-\delta$*,*

$$\mathbb{E}[\{\mathcal{T}\hat{b}_V\}^2(H)] \leq c\{1/\lambda + \lambda\} \max(1, C_{\mathcal{Q}}, C_{\Xi}) \sqrt{\frac{\ln(|\mathcal{Q}||\Xi|c/\delta)}{n}},$$

where c *is some universal constant.*

Theorem 6 (Finite sample property of $\hat{J}_{\mathrm{VM}}$)**.** *Set* $\lambda > 0$*. Suppose (5a), (5b), (5c) any element in* $q \in \mathcal{Q}$ *that satisfies* $\mathbb{E}[\{\mathcal{T}^S(q)\}(\bar{S}) \mid H] = 0$ *also satisfies* $\mathcal{T}^S(q)(\bar{S}) = 0$*. (5d) the overlap* $\mu(Z,O,A) < \infty$ *and any element in* $q \in \mathcal{Q}$ *that satisfies* $\mathcal{T}^S(q)(\bar{S}) = 0$ *also satisfies* $\mathcal{T}^S(q)(\bar{S}^\diamond) = 0$ *where* $\bar{S}^\diamond \sim d_{\pi^e}(\bar{s})$*. With probability* $1-\delta$*, we have*

$$|J(\pi^e) - \hat{J}_{\mathrm{VM}}| \leq c(1-\gamma)^{-2}(1/\lambda + \lambda) \max(1, C_{\mathcal{Q}}, C_{\Xi}) \mathrm{IV}_1(\mathcal{Q}) \mathrm{Dr}_{\mathcal{Q}}[d_{\pi^e}, P_{\pi^b}] \sqrt{\frac{\ln(|\mathcal{Q}||\Xi|c/\delta)}{n}}, \tag{12}$$

where

$$\mathrm{IV}_1^2(\mathcal{Q}) := \sup_{\{q \in \mathcal{Q}; \mathbb{E}[\{\mathcal{T}(q)(H)\}^2] \neq 0\}} \frac{\mathbb{E}[\{\mathcal{T}^{\mathcal{S}}(q)(\bar{S})\}^2]}{\mathbb{E}[\{\mathcal{T}(q)(H)\}^2]}, \tag{13}$$

$$\mathrm{Dr}_{\mathcal{Q}}^2[d_{\pi^e}, P_{\pi^b}] := \sup_{\{q \in \mathcal{Q}; \mathbb{E}_{\bar{s} \sim P_{\pi^b}}[\{\mathcal{T}^{\mathcal{S}}(q)(\bar{s})\}^2] \neq 0\}} \frac{\mathbb{E}_{s \sim d^{\pi^e}}[\{\mathcal{T}^{\mathcal{S}}(q)(\bar{s})\}^2]}{\mathbb{E}_{\bar{s} \sim P_{\pi^b}}[\{\mathcal{T}^{\mathcal{S}}(q)(\bar{s})\}^2]}. \tag{14}$$

D Examples

D.1 Tabular POMDPs

We have seen that in Lemma 2, $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{S}_b, \mathbf{H})) = |\mathcal{S}_b|$ and $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{F} \mid \mathbf{S}_b)) = |\mathcal{S}_b|$ are sufficient conditions for the identification in the tabular setting. The following theorem show that the abovementioned two conditions are equivalent to $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{F}, \mathbf{H})) = |\mathcal{S}_b|$.

Lemma 3. $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{S}_b, \mathbf{H})) = |\mathcal{S}_b|$ *and* $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{F} \mid \mathbf{S}_b)) = |\mathcal{S}_b|$ *holds if and only if* $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{F}, \mathbf{H})) = |\mathcal{S}_b|$*.*

We again make a few remarks. First, $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{F}, \mathbf{H})) = |\mathcal{S}|$ is often imposed to model HMMs and POMDPs [HKZ12, BSG11]. Here, our condition $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{F}, \mathbf{H})) = |\mathcal{S}_b|$ is weaker than this assumption. We discuss the connection to the aforementioned works in Section F. Second, in the literature of online RL, $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{F}, \mathbf{S})) = |\mathcal{S}|$ is frequently imposed as well [JKKL20, LCSJ22] although they don't impose assumptions associated with the history proxy H. In confounded POMDPs, [NJ21, SUHJ22] imposed a closely-related assumption, namely, $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{F}, \mathbf{H}, a)) = |\mathcal{S}|$ for any $a \in \mathcal{A}$ where $\Pr_{\pi^b}(\mathbf{F}, \mathbf{H}, a)$ is a matrix whose (i,j)-th element is $\Pr_{\pi^b}(F = x_i, H = x'_j, A = a)$ $(\mathcal{F} = \{x_i\}, \mathcal{H} = \{x'_j\})$.

D.2 HSE-POMDPs and LQGs

In this section, we primarily emphasize the identification results in HSE-POMDPs. Extending these results to the final sample result is straightforward.

Refined identification theorem. First, we describe the population version of Theorem 3 as follows.

Theorem 7 (Refined identification theorem)**.** *Suppose (7a) $\mathcal{B}_V \cap \mathcal{Q} \neq \emptyset$, (7b) any element in $q \in \mathcal{Q}$ that satisfies $\mathbb{E}[\{\mathcal{T}^S(q)\}(S) \mid H] = 0$ also satisfies $\mathcal{T}^S(q)(S) = 0$. (7c) the overlap $\mu(O, A) < \infty$ and any element in $q \in \mathcal{Q}$ that satisfies $\mathcal{T}^S(q)(S) = 0$ also satisfies $\mathcal{T}^S(q)(S^\diamond) = 0$ where $S^\diamond \sim d_{\pi^e}(s)$. Under the above three conditions, for any $b_V \in \mathcal{B}_V \cap \mathcal{Q}$, we have*

$$J(\pi^e) = \mathbb{E}_{f\sim\nu_{\mathcal{F}}}[b_V(f)].$$

The proof is deferred to Section L

HSE-POMDPS. By using Theorem 7, we can obtain a useful identification formula when we set $\mathcal{Q}, \Xi$ to be linear models in HSE-POMDPs. We start with the definition.

Example 5 (HSE-POMDPs with linear models)**.** *Introduce features $\phi_{\mathcal{F}} : \mathcal{F} \to \mathbb{R}^{d_{\mathcal{F}}}, \phi_{\mathcal{S}} : \mathcal{S}_b \to \mathbb{R}^{d_{\mathcal{S}}}, \phi_{\mathcal{H}} : \mathcal{H} \to \mathbb{R}^{d_{\mathcal{H}}}$ such that $\|\phi_{\mathcal{F}}(\cdot)\| \leq 1, \|\phi_{\mathcal{S}}(\cdot)\| \leq 1, \|\phi_{\mathcal{H}}(\cdot)\| \leq 1$. Letting $\mathcal{Q}$ and Ξ be linear models, the existence of future-dependent value functions in $\mathcal{Q}$ is ensured as follows under certain conditions. Then, the existence of learnable future-dependent value functions in $\mathcal{Q}$, (7a) is automatically satisfied.*

Next, we provide sufficient conditions for the realizability (7a).

Lemma 4. *Suppose (LM1): $\mathbb{E}[\phi_{\mathcal{F}}(F) \mid S] = K_1\phi_{\mathcal{S}}(S)$ for some $K_1 \in \mathbb{R}^{d_{\mathcal{F}}\times d_{\mathcal{S}}}$, (LM2): $V^{\pi^e}(S)$ is linear in $\phi_{\mathcal{S}}(S)$, i.e., $V^{\pi^e}(S) \in \{w^\top\phi_{\mathcal{S}}(S) : w \in \mathbb{R}^{d_{\mathcal{S}}}, \|w\| \leq C_{\mathrm{LM}}\}$ for some $C_{\mathrm{LM}} \in \mathbb{R}$, (LM3) for any $b \in \mathbb{R}^{\mathcal{S}}$ such that $\|b\| \leq C_{\mathrm{LM}}$, there exists $a \in \mathbb{R}^{d_{\mathcal{F}}}, \|a\| \leq C_{\mathcal{Q}}$ such that $a^\top K_1\phi_{\mathcal{S}}(S) = b^\top\phi_{\mathcal{S}}(S)$. Then, future-dependent value functions exist and belong to $\mathcal{Q} = \{w^\top\phi_{\mathcal{F}}(\cdot) : w \in \mathbb{R}^{d_{\mathcal{F}}}, \|w\| \leq C_{\mathcal{Q}}\}$ for some $C_{\mathcal{Q}} \in \mathbb{R}$.*

The condition (LM1) requires the existence of a conditional mean embedding operator between $\mathcal{F}$ and $\mathcal{S}$. This assumption is widely used to model PSRs, which include POMDPs and HMMs [SHSF09, BGG13]. In addition, assumptions of this type are frequently imposed to model MDPs as well [ZLKB20, DJW20, CO20, HDL$^+$21]. (LM2) is realizablity on the latent state space. (LM3) says the information of the latent space is not lost on the observation space. The condition $\mathrm{rank}(K_1) = d_{\mathcal{S}}$ is a sufficient condition; then, we can take $C_{\mathcal{Q}} = C_{\mathrm{LM}}/\sigma_{\min}(K_1)$. In the tabular case, we set $\phi_{\mathcal{F}}, \phi_{\mathcal{S}}, \phi_{\mathcal{H}}$ be one-hot encoding vectors over $\mathcal{F}, \mathcal{S}_b, \mathcal{H}$, respectively. Here, we remark that S in $\phi_{\mathcal{S}}(S)$ is a random variable in the offline data; thus, $\phi_{\mathcal{S}}(\cdot)$ needs to be just defined on the support of the offline data. Hence, (LM3) is satisfied when $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{F} \mid \mathbf{S}_b)) = |\mathcal{S}_b|$.

Next, we see (7b) is satisfied as follows under certain conditions.

Lemma 5. *Suppose (LM1), (LM2), (LM4): $\mathbb{E}[\mu(O, A)\phi_{\mathcal{S}}(S') \mid S]$ is linear in $\phi_{\mathcal{S}}(S)$ and (LM5):*

$$\sup_{x\in\mathbb{R}^d, x^\top\mathbb{E}[\mathbb{E}[\phi_{\mathcal{S}}(S)|H]\mathbb{E}[\phi_{\mathcal{S}}(S)|H]^\top]x\neq 0} \frac{x^\top\mathbb{E}[\phi_{\mathcal{S}}(S)\phi_{\mathcal{S}}(S)^\top]x}{x^\top\mathbb{E}[\mathbb{E}[\phi_{\mathcal{S}}(S) \mid H]\mathbb{E}[\phi_{\mathcal{S}}(S) \mid H]^\top]x} < \infty, \tag{15}$$

hold. Then, (7b) is satisfied.

Condition (LM4) requires the existence of conditional mean embedding between S' and S under the distribution induced by a policy π^e. The condition (LM5) is satisfied when $\mathrm{rank}(\Pr_{\pi^b}(\mathbf{S}_b, \mathbf{H})) = |\mathcal{S}_b|$ in the tabular setting.

Combining Lemma 4 and Lemma 5 with Theorem 7, we obtain the following handy formula.

Lemma 6 (Formula with linear models in HSE-PODMDPs)**.** *Suppose (LM1)-(LM5), (LM6): there exists a matrix $K_2 \in \mathbb{R}^{d_{\mathcal{S}}\times d_{\mathcal{H}}}$ such that $\mathbb{E}[\phi_{\mathcal{S}}(S) \mid H] = K_2\phi_{\mathcal{H}}(H)$, (LM7): $\mu(Z, O, A) < \infty$ and*

$$\sup_{x\in\mathbb{R}^d, 0\neq x^\top\mathbb{E}_{s\sim P_{\pi^b}}[\phi_{\mathcal{S}}(s)\phi_{\mathcal{S}}(s)^\top]x} \frac{x^\top\mathbb{E}_{s\sim d_{\pi^e}}[\phi_{\mathcal{S}}(s)\phi_{\mathcal{S}}(s)^\top]x}{x^\top\mathbb{E}_{s\sim P_{\pi^b}}[\phi_{\mathcal{S}}(s)\phi_{\mathcal{S}}(s)^\top]x} < \infty, \tag{16}$$

hold. Then, we have

$$J(\pi^e) = \mathbb{E}_{f\sim\nu_{\mathcal{F}}}[\phi_{\mathcal{F}}(f)]^\top\mathbb{E}[\phi_{\mathcal{H}}(H)\{\phi_{\mathcal{F}}(F) - \gamma\mu(O, A)\phi_{\mathcal{F}}(F')\}^\top]^+\mathbb{E}[\mu(O, A)R\phi_{\mathcal{H}}(H)]. \tag{17}$$

We imposed two additional assumptions in Lemma 6. (LM6) is used to ensure the Bellman completeness assumption $\mathcal{T}\mathcal{Q} \subset \Xi$. (LM7) is similar to the overlap condition (7c) in linear models. It is characterized based on a relative condition number whose value is smaller than the density ratio. Similar assumptions are imposed in offline RL with fully observable MDPs as well [XCJ$^+$21, ZWB21, US21].

LQG. Finally, we extend our result to the case of LQG.

Example 1 (LQG)**.** *Linear Quadratic Gaussian (LQG) falls in the category of HSE-POMDPs. Suppose*

$$s_{t+1} = As_t + Ba_t + \epsilon_{1t}, \quad r_t = -s_t^\top Qs_t - a_t^\top Ra_t, \quad o_t = Cs_t + \epsilon_{2t}$$

where A, B, C, Q, R are matrices that parametrize models and ϵ_{1t} and ϵ_{2t} are Gaussian noises. Consider a linear evaluation policy $\pi^e(a_t \mid o_t, z_t) = \mathrm{I}(a_t = F[o_t, z_t^\top]^\top)$ for certain matrix F. Notice that linear policies are commonly used in LQG since the globally optimal policy is linear [Ber12]. Then, defining $\phi_{\mathcal{S}}(s) = (1, \{s \otimes s\}^\top)^\top$, $\phi_{\mathcal{F}}(f) = (1, \{f \otimes f\}^\top)^\top$ and $\phi_{\mathcal{H}}(h) = (1, \{h \otimes h\}^\top)^\top$, the following holds.

Lemma 7. *In LQG, (LM1),(LM2), (LM4) are satisfied. When C is left-invertible, (LM3) holds.*

Thus, what we additionally need to assume is only (LM5), (LM6) and (LM7) in LQG.

E Finite Horizon Off-Policy Evaluation

For completeness, we also consider estimation of $J_T(\pi^e) = \mathbb{E}_{\pi^e}[\sum_{k=0}^{T-1} \gamma^k R_k]$ when the horizon is finite and the system dynamics are nonstationary. We first define value and learnable future-dependent value functions following Section 3. Let $V_t^{\pi^e}(s) = \mathbb{E}_{\pi^e}[\sum_{k=t}^{\infty} \gamma^{k-t} R_k \mid S_t = s]$ denote the state value function.

Definition 7 (Future-dependent value functions)**.** *For $t \in [T-1]$, future-dependent value functions $\{g_V^{[t]}\}_{t=0}^{T-1}$ are defined as solutions to*

$$0 = \mathbb{E}[g_V^{[t]}(F) \mid S] - V_t^{\pi^e}(S)$$

and $g_V^{[T]} = 0$. We denote the set of $g_V^{[t]}$ by $\mathcal{G}_V^{[t]}$.

Definition 8 (Learnable future-dependent value functions)**.** *For $t \in [T-1]$, learnable future-dependent value functions $\{b_V^{[t]}\}_{t=0}^{T-1}$ are defined as solutions to*

$$\mathbb{E}[\mu(O,A)\{R + \gamma b_V^{[t+1]}(F')\} - b_V^{[t]}(F) \mid H]$$

where $b_V^{[T]} = 0$. We denote the set of $b_V^{[t]}$ by $\mathcal{B}_V^{[t]}$.

We define the following Bellman operator:

$$\mathcal{T}^{\mathcal{S},t} : \prod_{t=0}^{T-1} [\mathcal{F} \to \mathbb{R}] \ni \{q_t(\cdot)\} \mapsto \mathbb{E}[\mu(O,A)\{R + \gamma q_{t+1}(F')\} - q_t(F) \mid S = \cdot] \in [\mathcal{S} \to \mathbb{R}].$$

We again remark that while the conditional expectation of the offline data is not defined on the outside of $\mathcal{S}_b$ (the support of $\mathcal{S}$) above, we just set 0 on the outside of $\mathcal{S}_b$.

Here are the analogs statements of Theorem 7 in the finite horizon setting.

Lemma 8. *Future-dependent value functions are learnable future-dependent value functions.*

Theorem 8 (Identification for finite horizon OPE)**.** *Suppose for any $t \in [T-1]$, (8a) $\mathcal{B}_V^{[t]} \cap \mathcal{Q}_t \neq \emptyset$, (8b) for any $q \in \mathcal{Q}_t$ that satisfies $\mathbb{E}[\{\mathcal{T}^{\mathcal{S},t}(q)\}(S) \mid H] = 0$ also satisfies $\{\mathcal{T}^{\mathcal{S},t}(q)\}(S) = 0$, (8c) overlap $\mu(O,A) < \infty$ and for any $q \in \mathcal{Q}_t$ that satisfies $\{\mathcal{T}^{\mathcal{S},t}(q)\}(S) = 0$ also satisfies $\{\mathcal{T}^{\mathcal{S},t}(q)\}(S_t^\diamond) = 0$ where $S_t^\diamond \sim d_t^{\pi^e}(\cdot)$. Then, for any $b_V^{[0]} \in \mathcal{B}_V^{[0]} \cap \mathcal{Q}_0$, we have*

$$J_T(\pi^e) = \mathbb{E}_{f \sim \nu_{\mathcal{F}}}[b_V^{[0]}(f)].$$

Here, (8a), (8b), (8c) correspond to (7a), (7b), (7c), respectively. Lemma 8 is often useful to ensure (8a).

In the tabular setting, when we have $\mathrm{rank}(\Pr_{\pi_b}(\mathbf{F}, \mathbf{H})) = |\mathcal{S}_b|$, conditions (a) and (b) are satisfied. This is the same condition imposed in Example 1. When we use linear models for $\mathcal{Q}_t$ and Ξ_t, we have the following corollary. This is the finite-horizon version of Lemma 6.

Corollary 1 (Formula with linear models in HSE-POMDPs)**.** *Suppose (LM1), (LM2f) $V^{\pi^e}_t(S)$ is linear in $\phi_{\mathcal{S}}(S)$ for any $t \in [T-1]$, (LM3), (LM4),(LM5), (LM6). Then under the overlap $\mathrm{Dr}_{\mathcal{Q}}(d^{\pi^e}_t, P_{\pi^b}) < \infty$ and $\mu(O, A) < \infty$ for any $t \in [T-1]$. Starting from $\theta_T = 0$, we recursively define*

$$\theta_t = \mathbb{E}[\phi_{\mathcal{H}}^\top(H)\phi_{\mathcal{F}}(F)]^{+}\mathbb{E}[\mu(O, A)\phi_{\mathcal{H}}(H)\{R + \gamma\theta_{t+1}^\top\phi_{\mathcal{F}}(F')\}].$$

Then, $J_T(\pi^e) = \mathbb{E}_{f\sim\nu_{\mathcal{F}}}[\theta_0^\top\phi_{\mathcal{F}}(f)]$.

Remark 7 (Comparison to SIS)**.** *When we have finite samples, the estimator is defined as in Section 4. Then, we can obtain the finite sample result as in Section 5. In this case, we can again possibly circumvent the curse of horizon. The error scales with the marginal ratio $\max_{t\in[T-1]}\max_{s\in\mathcal{S}}(d^{\pi^e}_t(s)/P_{\pi_b}(s))^{1/2}$. Compared to SIS, the finite sample error does not directly incur the exponential dependence on the horizon.*

F Modeling of System Dynamics

We have so far discussed how to estimate cumulative rewards under evaluation policies. In the literature on POMDPs [SBS+10, BSG11, BGG13, KJS15], we are often interested in learning of system dynamics. In this section, we discuss how our methods are extended to achieve this goal. We ignore rewards in this section. We assume policies are memory-less.

F.1 Tabular Setting

Here, let $S^\diamond_0, O^\diamond_0, A^\diamond_0, \cdots$ be random variables under a memory-less evaluation policy $\pi^e : \mathcal{O} \to \Delta(\mathcal{A})$. Following [SBS+10, BSG11, HKZ12], we consider two common estimands:

$$\mathrm{Pr}_{\pi^e}(o_0, a_0, \cdots, o_{T-1}, a_{T-1}) := \mathrm{Pr}_{\pi^e}(O^\diamond_0 = o_0, A^\diamond_0 = a_0, O^\diamond_1 = o_1, \cdots), \tag{18}$$

$$\mathrm{Pr}_{\pi^e}(\mathbf{O}_T \mid o_0, a_0, \cdots, a_{T-1}) := \{\mathrm{Pr}_{\pi^e}(O^\diamond_T = x_i \mid O^\diamond_0 = o_0, \cdots, A^\diamond_{T-1} = a_{T-1})\}_{i=1}^{|\mathcal{O}|}, \tag{19}$$

given a sequence $o_0 \in \mathcal{O}, \cdots, a_{T-1} \in \mathcal{A}$. Our goal is to estimate (18) and (19) from the offline data. To simplify the discussion, we first consider the tabular case. If we can model a $|\mathcal{O}|$-dimensional vector $\mathrm{Pr}_{\pi^e}(o_0, \cdots, a_{T-1}, \mathbf{O}_T) \in \mathbb{R}^{|\mathcal{O}|}$ where the entry indexed by $x_i \in \mathbf{O}$ is $\mathrm{Pr}_{\pi^e}(o_0, \cdots, a_{T-1}, x_i)$, the latter estimand is computed by normalizing a vector, i.e., dividing it over the sum of all elements. Therefore, we have $\mathrm{Pr}_{\pi^e}(\mathbf{O}_T \mid o_0, a_0, \cdots, a_{T-1}) \propto \mathrm{Pr}_{\pi^e}(o_0, \cdots, a_{T-1}, \mathbf{O}_T)$. Hereafter, we consider modeling $\mathrm{Pr}_{\pi^e}(o_0, \cdots, a_{T-1}, \mathbf{O}_T)$ instead of $\mathrm{Pr}_{\pi^e}(\mathbf{O}_T \mid o_0, \cdots, a_{T-1})$.

To identify estimands without suffering from the curse of horizon, we would like to apply our proposal in the previous sections. Toward that end, we set rewards as the product of indicator functions

$$\mathrm{Pr}_{\pi^e}(o_0, a_0, \cdots, o_{T-1}, a_{T-1}) = \mathbb{E}\left(\prod_{k=0}^{T-1}\mathrm{I}(O^\diamond_t = o_t, A^\diamond_t = a_t)\right).$$

Since this is a product but not a summation, we cannot directly use our existing results. Nevertheless, the identification strategy is similar.

We first introduce learnable future-dependent value functions. These are analogs of Definition 2 tailored to the modeling of system dynamics. Let $\Xi \subset [\mathcal{H} \to \mathbb{R}]$ and $\mathcal{Q} \subset [\mathcal{F} \to \mathbb{R}]$. Below, we fix $o_0 \in \mathcal{O}, \cdots, a_{T-1} \in \mathcal{A}$.

Definition 9 (Learnable future-dependent value functions for modeling dynamics)**.** *Learnable future-dependent value functions $\{b^{[t]}_D\}_{t=0}^{T-1}$ where $b^{[t]}_D : \mathcal{F} \to \mathbb{R}$, for joint observational probabilities are defined as solutions to*

$$0 \le t \le T-1; \mathbb{E}[b^{[t]}_D(F) - \mathrm{I}(O = o_t, A = a_t)\mu(O, A)b^{[t+1]}_D(F') \mid H] = 0,$$
$$\mathbb{E}[b^{[T]}_D(F) - 1 \mid H] = 0.$$

We denote the set of solutions $b^{[t]}_D$ by $\mathcal{B}^{[t]}_D$. Learnable future-dependent value functions $\{b^{[t]}_P\}_{t=0}^{T-1}$ where $b^{[t]}_P : \mathcal{F} \to \mathbb{R}^{|\mathcal{O}|}$ for conditional observational probabilities are defined as solutions to

$$0 \le t \le T-1; \mathbb{E}[b^{[t]}_P(F) - \mathrm{I}(O = o_t, A = a_t)\mu(O, A)b^{[t+1]}_P(F') \mid H] = \mathbf{0}_{|\mathcal{O}|},$$
$$\mathbb{E}[b^{[T]}_P(F) - \phi_{\mathcal{O}}(O) \mid H] = \mathbf{0}_{|\mathcal{O}|},$$

where $\phi_{\mathcal{O}}(\cdot)$ is a $|\mathcal{O}|$-dimensional one-hot encoding vector over $\mathcal{O}$ and $\mathbf{0}_{|\mathcal{O}|}$ is a $|\mathcal{O}|$-dimensional vector consisting of 0. We denote the set of solutions $b_P^{[t]}$ by $\mathcal{B}_P^{[t]}$.

Next, we define the Bellman operator.

Definition 10 (Bellman operators for modeling systems)**.**

$$\mathcal{T}_t^{\mathcal{S}}:\prod_{t=0}^{T-1}[\mathcal{F}\to\mathbb{R}]\ni\{q_t(\cdot)\}\mapsto\mathbb{E}[q_t(F)-\mathrm{I}(O=o_t,A=a_t)\mu(O,A)q_{t+1}(\mathcal{F}')\mid S=\cdot]\in[\mathcal{S}\to\mathbb{R}].$$

Following Theorem 7, we can identify estimands. Here, let $\grave{d}_t^{\pi}(\cdot)\in\Delta(\mathcal{S})$ be a probability density function of $S_t^{\diamond}$ conditional on $O_0^{\diamond}=o_0,\cdots,A_T^{\diamond}=a_{T-1}$.

Theorem 9 (Identification of joint probabilities)**.** *Suppose (9a) $\mathcal{B}_D^{[t]}\cap\mathcal{Q}_t\neq\emptyset$, (9b) for any $q\in\mathcal{Q}_t$ that satisfies $\mathbb{E}[(\mathcal{T}_t^{\mathcal{S}}q)(S)\mid H]=0$ also satisfies $(\mathcal{T}_t^{\mathcal{S}}q)(S)=0$,(9c) for any $q\in\mathcal{Q}_t$ that satisfies $(\mathcal{T}_t^{\mathcal{S}}q)(S)=0$ also satisfies $(\mathcal{T}_t^{\mathcal{S}}q)(\grave{S}_t)=0$ where $\grave{S}_t\sim\grave{d}_t^{\pi}(\cdot)$ and $\mu(O,A)<\infty$. Then, for any $b_D^{[0]}\cap\mathcal{Q}_0\in\mathcal{B}_D^{[0]}$, we have*

$$\mathrm{Pr}_{\pi^e}(o_0,a_0\cdots,o_{T-1},a_{T-1})=\mathbb{E}_{f\sim\nu_{\mathcal{F}}}[b_D^{[0]}(f)].$$

Theorem 10 (Identification of conditional probabilities)**.** *Suppose (10a) $\mathcal{B}_P^{[t]}\cap\mathcal{Q}_t\neq\emptyset$, (10b) for any $q\in\mathcal{Q}_t$ that satisfies $\mathbb{E}[\{\mathcal{T}_t^{\mathcal{S}}(q)\}(S)\mid H]=0$ also satisfies $(\mathcal{T}_t^{\mathcal{S}}q)(S)=0$, (10c) for any $q\in\mathcal{Q}_t$ that satisfies $(\mathcal{T}_t^{\mathcal{S}}q)(S)=0$ also satisfies $(\mathcal{T}_t^{\mathcal{S}}q)(\grave{S}_t)$ where $\grave{S}_t\sim\grave{d}_t^{\pi}(\cdot)$ and $\mu(O,A)<\infty$. Then, for any $b_P^{[0]}\cap\mathcal{Q}_0\in\mathcal{B}_P^{[0]}$, we have*

$$\mathrm{Pr}_{\pi^e}(o_0,a_0\cdots,o_{T-1},a_{T-1},\mathbf{O}_T)=\mathbb{E}_{f\sim\nu_{\mathcal{F}}}[b_P^{[0]}(f)].$$

The following corollary is an immediate application of Theorem 9 and Theorem 10.

Corollary 2 (Tabular Models)**.** *Let $\phi_{\mathcal{F}}(\cdot),\phi_{\mathcal{H}}(\cdot)$ be one-hot encoding vectors over $\mathcal{F}$ and $\mathcal{H}$, respectively. Suppose $\mathrm{rank}(\mathrm{Pr}(\mathbf{F},\mathbf{H}))=|\mathcal{S}_b|$ and $\grave{d}_t^{\pi^e}(\cdot)/P_{\pi^b}(\cdot)<\infty$ for any $t\in[T-1]$ where $P_{\pi^b}(\cdot)$ is a pdf of $\mathcal{S}$ in the offline data. Then, we have*

$$\mathrm{Pr}_{\pi^e}(o_0,a_0,\cdots,o_{T-1},a_{T-1})=\mathrm{Pr}_{\pi^b}(\mathbf{H})^{\top}B^{+}\left\{\prod_{t=T-1}^{0}D_tB^{+}\right\}C,\tag{20}$$

$$\mathrm{Pr}_{\pi^e}(\mathbf{O}_T\mid o_0,a_0,\cdots,o_{T-1},a_{T-1})\propto\mathrm{Pr}_{\pi^b}(\mathbf{O},\mathbf{H})B^{+}\left\{\prod_{t=T-1}^{0}D_tB^{+}\right\}C,\tag{21}$$

where $B=\mathrm{Pr}_{\pi^b}(\mathbf{F},\mathbf{H}),D_t=\mathbb{E}[\mathrm{I}(O=o_t,A=a_t)\mu(O,A)\phi_{\mathcal{F}}(F')\phi_{\mathcal{H}}^{\top}(H)],C=\mathrm{Pr}_{\nu_{\mathcal{F}}}(\mathbf{F})$.

In particular, when behavior policies are uniform policies and evaluation policies are atomic i.e., $\pi_t^e(a)=\mathrm{I}(a=a_t)$ for some a_t and any t, we have $D_t=\mathrm{Pr}_{\pi^b}(O=o_t,\mathbf{F}',\mathbf{H}\mid A=a_t)$. In addition, the rank assumption is reduced to $\mathrm{rank}(\mathrm{Pr}_{\pi^b}(\mathbf{F},\mathbf{H}))=|\mathcal{S}_b|$.

Linear models. Next, we consider cases where $\mathcal{Q}_t,\Xi_t$ are linear models as in Example 2. We first define value functions:

$$V_{D,[t]}^{\pi^e}(\cdot)=\mathrm{Pr}_{\pi^e}(O_t^{\diamond}=o_t,A_t^{\diamond}=a_t,\cdots,O_T^{\diamond}=O_T,A_T^{\diamond}=a_T\mid S_t^{\diamond}=\cdot),$$
$$V_{P,[t]}^{\pi^e}(\cdot)=\{\mathrm{Pr}_{\pi^e}(O_t^{\diamond}=o_t,A_t^{\diamond}=a_t,\cdots,O_T^{\diamond}=x_i\mid S_t^{\diamond}=\cdot)\}_{i=1}^{|\mathcal{O}|}.$$

Then, we can obtain the following formula as in Lemma 6. Here, $\mathrm{Dr}_{\mathcal{Q}}(\grave{d}_t^{\pi^e},P_{\pi^b})$ is the condition number in Lemma 6.

Corollary 3 (Formula with linear models in HSE-POMDPs)**.** *Suppose (LM1), (LM2D) $V_{D,[t]}^{\pi^e}(S)$ is linear in $\phi_{\mathcal{S}}(S)$, (LM3), (LM4D) $\mathbb{E}[\mu(O,A)\mathrm{I}(O=o_t,A=a_t)\phi_{\mathcal{F}}(F')\mid S],\mathbb{E}[1\mid S]$ is linear in $\phi_{\mathcal{S}}(S)$, (LM5) and (LM6D) $\mathrm{Dr}_{\mathcal{Q}}(\grave{d}_t^{\pi^e},P_{\pi^b})<\infty$. Then, we have*

$$\mathrm{Pr}_{\pi^e}(o_0,a_0,\cdots,o_{T-1},a_{T-1})=\mathbb{E}[\phi_{\mathcal{H}}(H)]^{\top}B^{+}\left\{\prod_{t=T-1}^{0}D_tB^{+}\right\}C\tag{22}$$

Algorithm 3 Minimax Modeling of Dynamics on POMDPs

```
Require: Dataset $\mathcal{D}$, function classes $\mathcal{Q} \subset [\mathcal{F} \to \mathbb{R}], \Xi \subset [\mathcal{H} \to \mathbb{R}]$, hyperparameter $\lambda \geq 0$, Horizon $T$
Set
    $$\hat{b}^{[T]}_D = \operatorname*{argmin}_{q \in \mathcal{Q}} \max_{\xi \in \Xi} \mathbb{E}_{\mathcal{D}} \left[ \{1 - q(F)\}\xi(H) - 0.5\lambda\xi^2(H) \right].$$
for $t = T - 1$ do
    $$\hat{b}^{[t]}_D = \operatorname*{argmin}_{q \in \mathcal{Q}} \max_{\xi \in \Xi} \mathbb{E}_{\mathcal{D}} \left[ \{\mathrm{I}(O = o_t, A = a_t)\mu(O, A)\hat{b}^{[t+1]}_D(F') - q(F)\}\xi(H) - 0.5\lambda\xi^2(H) \right]. \tag{24}$$
    $t \leftarrow t - 1$
end for
return $\hat{J}_{\mathrm{VM}} = \mathbb{E}_{f \sim \nu_{\mathcal{F}}}[\hat{b}^{[0]}_D(f)]$
```

where

$$B = \mathbb{E}[\phi_{\mathcal{F}}(F)\phi_{\mathcal{H}}(H)^\top], D_t = \mathbb{E}[\mathrm{I}(O = o_t, A = a_t)\mu(O, A)\phi_{\mathcal{F}}(F)\phi_{\mathcal{H}}(H)^\top], C = \mathbb{E}_{f \sim \nu_{\mathcal{F}}}[\phi_{\mathcal{F}}(f)].$$

Suppose (LM1), (LM2P) $V^{\pi^e}_{P,[t]}(\cdot)$ *is linear in* $\phi_{\mathcal{S}}(\cdot)$*, (LM3), (LM4P)* $\mathbb{E}[\mu(O, A)\mathrm{I}(O_t = o_t, A_t = a_t)\phi_{\mathcal{F}}(F') \mid S], \mathbb{E}[\phi_{\mathcal{O}}(\mathcal{O}) \mid S]$ *is linear in* $\phi_{\mathcal{S}}(S)$*, (LM5) and (LM6P)* $\mathrm{Dr}_{\mathcal{Q}}(\grave{d}^{\pi^e}_t, P_{\pi^b}) < \infty$*. Then,*

$$\mathrm{Pr}_{\pi^e}(\mathbf{O}_T \mid o_0, a_0, \cdots, o_{T-1}, a_{T-1}) \propto \mathbb{E}[\phi_{\mathcal{O}}(O)\phi^\top_{\mathcal{H}}(H)]B^+ \left\{ \prod_{t=T-1}^{0} D_t B^+ \right\} C. \tag{23}$$

When behavior policies are uniform, the formulas in (20) and (22) are essentially the same to those obtained via spectral learning [HKZ12, BSG11]. We emphasize that (20)–(23) appear to be novel to the literature since we consider the offline setting.

General function approximation. Finally, we introduce an algorithm to estimate joint probabilities in the tabular case with general function approximation, summarized in Algorithm 3. The conditional probabilities can be similarly estimated. We remark that function approximation is extremely useful in large-scale RL problems.

F.2 Non-Tabular Setting

We have so far focused on the tabular case. In this section, we consider the non-tabular case. Our goal is to estimate joint probabilities $\mathrm{Pr}_{\pi^e}(o_0, a_0, \cdots, a_{T-1})$ in (18) and

$$\mathbb{E}_{\pi^e}[\phi_{\mathcal{O}}(O_T) \mid o_0, a_0, \cdots, a_{T-1}]$$

where $\phi_{\mathcal{O}} : \mathcal{O} \to \mathbb{R}$. When $\phi_{\mathcal{O}}(\cdot)$ is a one-hot encoding vector, this is equivalent to estimating $\mathrm{Pr}_{\pi^e}(\mathbf{O}_T \mid o_0, a_0, \cdots, a_{T-1})$ in (19).

In the non-tabular case, Theorem 9 and Theorem 10 still hold by defining learnable future-dependent value functions $b^{[t]} : \mathcal{F} \to \mathbb{R}$ as solutions to

$$\mathbb{E}[b^{[t]}(F) \mid H] = \mathbb{E}[\mu(O, A)b^{[t]}(F') \mid H, O = o_t, A = a_t]\mathrm{Pr}_{\pi^b}(O = o_t, A = a_t).$$

where $b^{[t]}$ is either $b^{[t]}_D$ or $b^{[t]}_P$. Then, Corollary 3 holds by just replacing D_t with

$$\mathbb{E}[\mu(O, A)\phi_{\mathcal{F}}(F)\phi^\top_{\mathcal{H}}(H) \mid O = o_t, A = a_t]\mathrm{Pr}_{\pi^b}(O = o_t, A = a_t).$$

When we have a finite sample of data, we need to perform density estimation for $\mathrm{Pr}_{\pi^b}(O = o_t, A = a_t)$. This practically leads to instability of estimators. However, when our goal is just to estimate $\mathbb{E}[\phi_{\mathcal{O}}(O) \mid o_0, a_0, \cdots, a_{T-1}]$ up to some scaling constant as in [SBS+10], we can ignore $\Pr_{\pi^b}(O = o_t, A = a_t)$. Then, we obtain the following formula:

$$\mathbb{E}_{\pi^e}[\phi_{\mathcal{O}}(O_T) \mid o_0, a_0, \cdots, a_{T-1}] \propto \mathbb{E}[\phi_{\mathcal{O}}(O)\phi_{\mathcal{H}}^\top(H)]B^+ \left\{ \prod_{t=T-1}^{0} \mathbb{E}[\mu(O,A)\phi_{\mathcal{F}}(F)\phi_{\mathcal{H}}^\top(H) \mid O = o_t, A = a_t]B^+ \right\} C. \tag{25}$$

This formula is known in HMMs [SBS+10] and in POMDPs [BGG13] where behavior policies are open-loop.

G Connection with Literature on Spectral Learning

We discuss the connection with previous literature in detail. [SBS+10, HKZ12] consider the modeling of HMMs with no action. In this case, our task is to model

$$\Pr(o_0, \cdots, o_{T-1}) := \Pr(O_0 = o_0, \cdots, O_{T-1} = o_{T-1})$$

and the predictive distribution:

$$\Pr(\mathbf{O}_T \mid o_0, \cdots, o_{T-1}) := \{\Pr(O_T = x_i \mid O_0 = o_0, \cdots, O_{T-1} = o_{T-1})\}_{i=1}^{|\mathcal{O}|}.$$

In the tabular case, the Corollary 2 is reduced to

$$\Pr(o_0, \cdots, o_{T-1}) = \Pr(\mathbf{H})^\top \Pr(\mathbf{F}, \mathbf{H})^+ \left\{ \prod_{t=T-1}^{0} \Pr(O = o_t, \mathbf{F}', \mathbf{H})\Pr(\mathbf{F}, \mathbf{H})^+ \right\} \Pr_{\nu_{\mathcal{F}}}(\mathbf{F}_0),$$

$$\Pr(\mathbf{O}_T \mid o_0, \cdots, o_{T-1}) \propto \Pr(\mathbf{O}, \mathbf{H})\Pr(\mathbf{F}, \mathbf{H})^+ \left\{ \prod_{t=T-1}^{0} \Pr(O = o_t, \mathbf{F}', \mathbf{H})\Pr(\mathbf{F}, \mathbf{H})^+ \right\} \Pr_{\nu_{\mathcal{F}}}(\mathbf{F}_0).$$

This formula is reduced to the one in [HKZ12] when $\mathcal{F} = \mathcal{H}$ and $\Pr(\mathbf{H}) = \Pr_{\nu_{\mathcal{F}}}(\mathbf{F}_0)$ (stationarity). Here, the offline data consists of three random variables $\{O_{-1}, O_0, O_1\}$. In this case, the above formulae are

$$\Pr(o_0, \cdots, o_{T-1}) = \Pr(\mathbf{O}_{-1})^\top \Pr(\mathbf{O}_0, \mathbf{O}_{-1})^+ \left\{ \prod_{t=T-1}^{0} \Pr(O_0 = o_t, \mathbf{O}_1, \mathbf{O}_{-1})\Pr(\mathbf{O}_0, \mathbf{O}_{-1})^+ \right\} \Pr(\mathbf{O}_{-1}),$$

$$\Pr(\mathbf{O}_T \mid o_0, \cdots, o_{T-1}) \propto \Pr(\mathbf{O}_0, \mathbf{O}_{-1})\Pr(\mathbf{O}_0, \mathbf{O}_{-1})^+ \left\{ \prod_{t=T-1}^{0} \Pr(O_0 = o_t, \mathbf{O}_1, \mathbf{O}_{-1})\Pr(\mathbf{O}_0, \mathbf{O}_{-1})^+ \right\} \Pr(\mathbf{O}_{-1}).$$

Next, we consider the case when we use linear models to estimate

$$\mathbb{E}[\phi_{\mathcal{O}}(O_T) \mid o_0, \cdots, o_{T-1}]$$

up to some constant scaling. When there are no actions, the formula (25) reduces to

$$\mathbb{E}[\phi_{\mathcal{O}}(O)\phi_{\mathcal{H}}^\top(H)]\mathbb{E}[\phi_{\mathcal{F}}(F)\phi_{\mathcal{H}}^\top(H)]^+ \left\{\textstyle\prod_{t=T-1}^{0} \mathbb{E}[\phi_{\mathcal{F}}(F')\phi_{\mathcal{H}}^\top(H) \mid O = o_t]\mathbb{E}[\phi_{\mathcal{F}}(F)\phi_{\mathcal{H}}^\top(H)]^+\right\} \mathbb{E}_{f\sim\nu_{\mathcal{F}}}[\phi_{\mathcal{F}}(f)].$$

When $\mathcal{F} = \mathcal{H}$, the pdf of O_{-1} is the same as $\nu_{\mathcal{F}}(\cdot)$ and the offline data consists of three random variables $\{O_{-1}, O_0, O_1\}$, the above is reduced to the one in [SBS+10] as follows:

$$\mathbb{E}[\phi_{\mathcal{O}}(O_0)\phi_{\mathcal{O}}^\top(O_{-1})]\mathbb{E}[\phi_{\mathcal{O}}(O_0)\phi_{\mathcal{O}}^\top(O_{-1})]^+ \left\{\textstyle\prod_{t=T-1}^{0} \mathbb{E}[\phi_{\mathcal{O}}(O_1)\phi_{\mathcal{O}}^\top(O_{-1}) \mid O_0 = o_t]\mathbb{E}[\phi_{\mathcal{O}}(O_0)\phi_{\mathcal{O}}^\top(O_{-1})]^+\right\} \mathbb{E}[\phi_{\mathcal{O}}(O_{-1})].$$

H Omitted Experiment Details and Additional Results

This section provides additional implementation details and ablation results of the synthetic experiment.

H.1 Sequential Importance Sampling (SIS)

We compare SIS [Pre00] as a baseline estimator. SIS is a non-parametric approach, which corrects the distribution shift between the behavior and evaluation policies by applying importance sampling as follows.

$$\hat{J}_{\mathrm{SIS}}(\pi^e; \mathcal{D}) := \mathbb{E}_{\mathcal{D}}\left[\sum_{t=1}^{\infty} \gamma^{t-1} \left(\prod_{t'=1}^{t} \frac{\pi^e(O_{t'} \mid A_{t'})}{\pi^b(O_{t'} \mid A_{t'})}\right) R_t\right]$$

SIS enables unbiased estimation when absolute continuity (i.e., $\forall(O, A) \in \mathcal{O} \times \mathcal{A}, \pi^e(O|A) > 0 \to \pi^b(O|A) > 0$) holds. However, as the importance weight grows exponentially as t becomes large, SIS suffers from high variance [LLTZ18, UHJ20, SUHJ22]. We also empirically verify that SIS incurs high estimation error due to variance in the experimental results.

H.2 Evaluation Metrics

We use the same evaluation metrics with [SUHJ22]. Given the i.i.d. dataset $\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_m$, the values estimated on them $\hat{J}_1, \hat{J}_2, \cdots, \hat{J}_m$, and the true value J^{π^e}, we define the relative bias and relative MSE as follows:

$$\text{Bias}(\hat{J}; \mathcal{D}, \pi) := \left| \frac{1}{m} \sum_{j=1}^{m} \left(\frac{\hat{J}_i - J^{\pi^e}}{J^{\pi^e}} \right) \right|, \qquad \text{MSE}(\hat{J}; \mathcal{D}, \pi) := \frac{1}{m} \sum_{j=1}^{m} \left(\frac{\hat{J}_i - J^{\pi^e}}{J^{\pi^e}} \right)^2$$

We use the above metrics to compare the performance of SIS, the naive baseline, and our proposal.

H.3 CartPole Setting

Environment. Here, we describe the state, action, and reward settings of the CartPole environment. First, the 4-dimensional states of CartPole represent the position and velocity of the cart and the angle and angle velocity of the pole. The action space is $\{0, 1\}$, either pushing the cart to the left or right. To better distinguish the values among different policies, we used modified reward following [SUHJ22]. Specifically, we define the reward as

$$R = \frac{1}{2} \left(\left| 2.0 - \frac{x}{x_{\text{clip}}} \right| \cdot \left| 2.0 - \frac{\theta}{\theta_{\text{clip}}} \right| - 1.0 \right),$$

where x and θ are the positions of Cart and angle of Pole. x_{clip} and θ_{clip} are the thresholds such that the episode will terminate when either $|x| \geq x_{\text{clip}}$ or $|\theta| \geq \theta_{\text{clip}}$. Under this definition, we observe a larger reward when the cart is closer to the center, and the pole stands straight. We also set the discount factor $\gamma = 0.95$, and the values of the policies used in our experiment are somewhere between 20 and 40.

Estimator Implementation. We parametrize the value function $b_v(\cdot)$ of our proposal and the naive estimator with a two-layer neural network. The network uses a 100-dimensional hidden layer with ReLU as its activation function, and Adam [KB14] is its optimizer. Both the naive estimator and our proposal optimize $b_v(\cdot)$ with the loss function defined in Example 5, but the naive one replaces $\bar{F}$ and H with O. Specifically, the naive estimator takes 4-dimensional observation O as input. On the other side, our proposed estimator additionally inputs the concatenated vector of observation O and one-hot representation of A for several future steps (i.e., M_F) to consider. The convergence is based on the test loss evaluated on the test dataset, which is independent of the datasets to train value functions and estimate the policy value. Specifically, to find a global convergence point, we first run 20000 epochs with 10 gradient steps per each. Then, we report the results on the epoch which achieves the minimum loss. The convergence point is usually less than 10000 epochs.

For the adversarial function space Ξ of both methods, we use the following RBF kernel $K(x_i; x_2)$:

$$K(x_1; x_2) := \exp\left(-\frac{\|x_1 - x_2\|_2}{2\sigma^2} \right),$$

where $\|x_1 - x_2\|_2$ is the l2-distance between x_1 and x_2, and σ is a bandwidth hyperparameter. We use $\sigma = 1.0$ in the main text and provide ablation results with varying values of σ in the following section. Finally, the naive estimator uses O as the input of the kernel. The proposed method first predicts the latent spaces from the historical observations as $\hat{S} := f_{\text{LSTM}}(H)$ and then use $\hat{S}$ as the input of the kernel. $f_{\text{LSTM}}(H)$ is a bi-directional LSTM [CKPW18] with 10-dimensional hidden dimension. We train the LSTM with MSE loss in predicting the noisy state (i.e., O) with Adam [KB14] as its optimizer.

H.4 Additional Rsults

Figure 6 shows the experimental results in the case of using the behavior policy with $\epsilon = 0.1$. The result suggests that the proposed estimator reduces MSEs of the baseline estimators as observed in

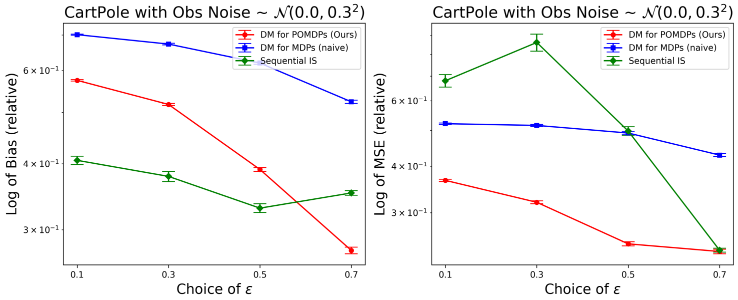


Figure 6: Logarithms of relative biases (left) and MSEs (right) of the proposed and the baseline estimators for various values of ϵ, which specify the evaluation policy. The confidence intervals are obtained through 100 Monte Carlo simulations.

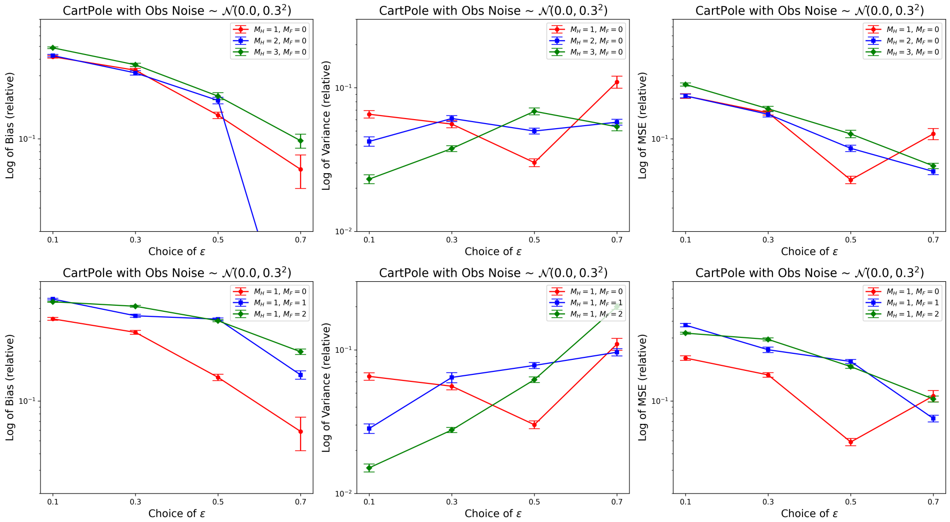


Figure 7: Logarithms of relative biases (left), variances (center), and MSEs (right) of the proposed estimator with varying lengths of history M_H (top) and varying lengths of future steps M_F (bottom). The x-axis corresponds to the varying values of ϵ of the evaluation policy, and the associated confidence interval is based on 20 simulations.

the $\epsilon = 0.3$ case in the main text. Note that experimental settings other than ϵ of the behavior policy are the same as those used in Section 6.

H.5 Ablation Results

Here, we provide ablation results with (1) varying choices of $\bar{F}$ and H, and (2) varying values of bandwidth hyperparameter σ of RKHSs. We provide the results with 20 random seeds in the following to conduct comprehensive ablations.

The first set of experiments aims to study how the use of history and future observations help improve the accuracy of value estimation. For this, we compare our proposed method with varying values of history length M_H and future length M_F. Figure 7 (top) shows the result of varying history lengths $M_H \in \{1, 2, 3\}$ with a fixed future length ($M_F = 0$). We observe that the performance of the proposed method does not change greatly with the choice of history lengths. This result suggests that 1-step history is almost sufficient to identify the latent state in our experimental settings. Next, we report the results with varying future lengths $M_F \in \{0, 1, 2\}$ with a fixed history length ($M_H = 1$) in Figure 7 (bottom). The result suggests that the increased number of future steps can slightly increase bias. We attribute this to the increased estimation difficulty of the value function due to the increase in the dimensionality of inputs. However, we should also note that future observations may help

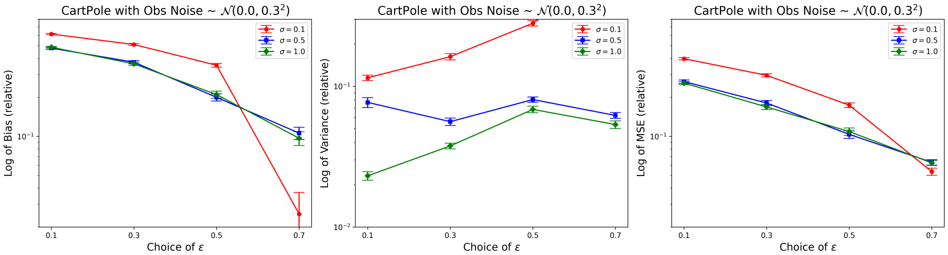


Figure 8: Logarithms of relative biases (left), variances (center), and MSEs (right) of the proposed estimator with varying bandwidth hyperparameter σ of RKHSs. The x-axis corresponds to the varying values of ϵ of the evaluation policy, and the associated confidence interval is based on 20 simulations.

improve the performance of the proposed method when the history itself is insufficient to identify the latent state.

The second set of experiments is to see how robust the proposed method is to the choice of hyperparameter σ. We thus vary the values of $\sigma \in [0.1, 0.5, 1.0]$, and report the result in Figure 8. The result shows that the estimation accuracy is almost the same between $\sigma = 1.0$ and $\sigma = 0.5$, suggesting that the proposed value learning method is robust to the change of bandwidth hyperparameter of RKHSs to some extent. On the other hand, we observe that a very small value (i.e., $\sigma = 0.1$) can increase the variance of estimation. Therefore, our recommendation is to avoid a (too) small value for the bandwidth hyperparameter σ.

I Proof of Section 3

We often use

$$F' \perp S, O, A \mid S' \quad \text{and} \quad (H) \perp (A, O, R, F') \mid S.$$

This is easily checked by graphical models Figure 4.

I.1 Proof of Lemma 1

From the Bellman equation, we have

$$\mathbb{E}[\mu(O,A)\{R + \gamma V^{\pi^e}(S')\} - V^{\pi^e}(S) \mid S] = 0.$$

Then, from the definition of future-dependent value functions,

$$\mathbb{E}[\mu(O,A)\{R + \gamma \mathbb{E}[g_V(F') \mid S']\} - \mathbb{E}[g_V(F) \mid S] \mid S] = 0.$$

Here, we use the stationarity assumption to ensure $\mathbb{E}[g_V(F') \mid S'] = V^{\pi^e}(S')$.

Next, by using $F' \perp S, O, A \mid S'$, we have

$$\mathbb{E}[\mu(O,A)\{R + \gamma g_V(F')\} - g_V(F) \mid S] = 0. \tag{26}$$

More specifically,

$$\begin{aligned}
&\mathbb{E}[\mu(O,A)\{R + \gamma g_V(F')\} - g_V(F) \mid S] \\
&= \mathbb{E}[\mu(O,A)\{R + \gamma \mathbb{E}[g_V(F') \mid S', S, O, A, Z]\} - \mathbb{E}[g_V(F) \mid S] \mid S] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Law of total expectation)} \\
&= \mathbb{E}[\mu(O,A)\{R + \gamma \mathbb{E}[g_V(F') \mid S']\} - \mathbb{E}[g_V(F) \mid S] \mid S] \qquad \text{(Use } F' \perp S, O, A \mid S'\text{)} \\
&= 0.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\mathbb{E}[\mu(O,A)\{R + \gamma g_V(F')\} - g_V(F) \mid H] \\
&= \mathbb{E}[\mathbb{E}[\mu(O,A)\{R + \gamma g_V(F')\} - g_V(F) \mid S, H] \mid H] \qquad \text{(Law of total expectation)} \\
&= \mathbb{E}[\mathbb{E}[\mu(O,A)\{R + \gamma g_V(F')\} - g_V(F) \mid S] \mid H] \qquad \text{(Use } R, O, A, F' \perp (H) \mid S\text{)} \\
&= 0. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(From (26))}
\end{aligned}$$

Thus, g_V is a learnable future-dependent value function.

I.2 Proof of Theorem 1

It follows from Theorem 7, which is an improved version of Theorem 1.

I.3 Proof of Lemma 2

The first statement is straightforward noting the equation is equal to solving

$$x^\top \mathrm{Pr}_{\pi^b}(\mathbf{F} \mid \mathbf{S}_b) = y$$

for x given y.

The second statement is straightforward noting it is equivalent to

$$x^\top \mathrm{Pr}_{\pi^b}(\mathbf{S}_b \mid \mathbf{H}) = 0$$

implies $x = 0$. This is satisfied when $\mathrm{rank}(\mathrm{Pr}_{\pi^b}(\mathbf{S}_b, \mathbf{H})) = |\mathcal{S}_b|$.

J Proof of Section 5

J.1 Proof of Theorem 2 and Theorem 3

By simple algebra, the estimator is written as

$$\hat{b}_V = \inf_{q\in\mathcal{Q}} \sup_{\xi\in\Xi} \mathbb{E}_{\mathcal{D}}[(\mathcal{Z}f)^2 - \{\mathcal{Z}f - \lambda\xi(H)\}^2]$$

where

$$\mathcal{Z}f = \mu(A,O)\{R + \gamma q(F')\} - q(F).$$

Noting this form similarly appears in the proof of [SUHJ22, Theorem 3], the following is similarly completed by [SUHJ22, Theorem 3]:

$$\mathbb{E}[\{\mathcal{T}(\hat{b}_V)\}^2(H)]^{1/2} = \tilde{O}\left(\max(1, C_{\mathcal{Q}}, C_{\Xi})\,(1/\lambda + \lambda)\sqrt{\frac{\ln(|\mathcal{G}||\Xi|/\delta)}{n}}\right).$$

using realizability and the Bellman completeness. Then, we have

$$\begin{aligned}
&(J(\pi^e) - \mathbb{E}_{f\sim\nu_{\mathcal{F}}}[\hat{b}_V(f)])^2 \\
&= \left\{(1-\gamma)^{-1}\mathbb{E}_{(\tilde{s})\sim d_{\pi^e}}\left[\mathbb{E}[\mu(A,O)R \mid S = \tilde{s}]\right] - \mathbb{E}_{f\sim\nu_{\mathcal{F}}}[\hat{b}_V(f)]\right\}^2 \\
&= \left\{(1-\gamma)^{-1}\mathbb{E}_{(\tilde{s})\sim d_{\pi^e}}\left[\mathbb{E}\left[\mu(A,O)\{R + \gamma\hat{b}_V(F')\} - \hat{b}_V(F) \mid S = \tilde{s}\right]\right]\right\}^2 && \text{(Use Lemma 9)} \\
&\leq (1-\gamma)^{-2}\mathbb{E}[\{\mathcal{T}^{\mathcal{S}}(\hat{b}_V)\}^2 s] && \text{(Jensen's inequality)} \\
&\leq (1-\gamma)^{-2}\mathbb{E}[\{\mathcal{T}(\hat{b}_V)\}^2(H)] \times \sup_{q\in\mathcal{Q}} \frac{\mathbb{E}_{s\sim d_{\pi^e}}[(\mathcal{T}^{\mathcal{S}}q)^2 s]}{\mathbb{E}[(\mathcal{T}q)^2(H)]}.
\end{aligned}$$

K Proof of Section C

In this section, most of the proof follows by slightly modifying the proof for memoryless policies. For completeness, we provide the proof of Theorem 4. As we did in the proof of Theorem 1, we prove the following stronger statement.

Theorem 11 (Refined identification theorem). *Suppose (7a) $\mathcal{B}_V \cap \mathcal{Q} \neq \emptyset$, (7b) any element in $q \in \mathcal{Q}$ that satisfies $\mathbb{E}[\{\mathcal{T}^S(q)\}(\bar{S}) \mid H] = 0$ also satisfies $\mathcal{T}^S(q)(\bar{S}) = 0$. (7c) the overlap $\mu(Z,O,A) < \infty$ and any element in $q \in \mathcal{Q}$ that satisfies $\mathcal{T}^S(q)(\bar{S}) = 0$ also satisfies $\mathcal{T}^S(q)(\bar{S}^\diamond) = 0$ where $S^\diamond \sim d_{\pi^e}(s)$. Under the above three conditions, for any $b_V \in \mathcal{B}_V \cap \mathcal{Q}$, we have*

$$J(\pi^e) = \mathbb{E}_{\bar{f}\sim\nu_{\bar{\mathcal{F}}}}[b_V(\bar{f})].$$

K.1 Proof of Theorem 11

Note for any $b_V \in \mathcal{B}_V \cap \mathcal{Q}$, we have

$$\begin{aligned}
0 &= \mathbb{E}\left[\mu(Z,A,O)\left(R+\gamma b_V(Z',F')\right) - b_V(Z,F) \mid H\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mu(Z,A,O)\left(R+\gamma b_V(Z',F')\right) - b_V(Z,F) \mid Z,S,H\right] \mid H\right] \quad \text{(Law of total expectation)} \\
&= \mathbb{E}\left[\mathbb{E}\left[\mu(Z,A,O)\left(R+\gamma b_V(Z',F')\right) - b_V(Z,F) \mid Z,S\right] \mid H\right].
\end{aligned}$$

In the last line, we use $(H \setminus Z) \perp (A,O,F') \mid Z,S$. From (7b), we have

$$\mathbb{E}[b_V(Z,F) \mid Z,S] = \mathbb{E}\left[\mu(Z,A,O)\left(R+\gamma b_V(Z',F')\right) \mid Z,S\right]. \tag{27}$$

Hence, $\mathcal{T}^{\mathcal{S}}(b_V)(S) = 0$. Then, from the overlap condition (7c),

$$\mathcal{T}^{\mathcal{S}}(b_V)(S^{\diamond}) = 0. \tag{28}$$

Finally, for any $b_V \in \mathcal{B}_V \cap \mathcal{Q}$, we have

$$\begin{aligned}
&(J(\pi^e) - \mathbb{E}_{\bar f \sim \nu_{\bar{\mathcal{F}}}}[b_V(\bar f)]) \\
&= (1-\gamma)^{-1}\mathbb{E}_{(\tilde s)\sim d_{\pi^e}}\left[\mathbb{E}[\mu(Z,A,O)R \mid \bar S = \tilde s]\right] - \mathbb{E}_{\bar f \sim \nu_{\bar{\mathcal{F}}}}[b_V(\bar f)] \\
&= (1-\gamma)^{-1}\mathbb{E}_{(\tilde s)\sim d_{\pi^e}}\left[\mathbb{E}\left[\mu(Z,A,O)\{R+\gamma b_V(Z',F')\} - b_V(Z,F) \mid \bar S = \tilde s\right]\right] \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{(Use Lemma 9)} \\
&= 0. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{Use } \mathcal{T}^{\mathcal{S}}(S^{\diamond}) = 0.)
\end{aligned}$$

From the first line to the second line, we use

$$\begin{aligned}
J(\pi^e) &= (1-\gamma)^{-1}\int d_{\pi^e}(z,s)r(s,a)\pi^e(a \mid z,o)\mathrm{d}(z,s) \\
&= (1-\gamma)^{-1}\mathbb{E}_{(\tilde s)\sim d_{\pi^e}}\left[\mathbb{E}[\mu(Z,A,O)R \mid \bar S = \tilde s]\right].
\end{aligned}$$

L Proof of Section D

L.1 Proof of Theorem 7

Note for any $b_V \in \mathcal{B}_V \cap \mathcal{Q}$, we have

$$\begin{aligned}
0 &= \mathbb{E}\left[\mu(A,O)\left(R+\gamma b_V(F')\right) - b_V(F) \mid H\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\mu(A,O)\left(R+\gamma b_V(F')\right) - b_V(F) \mid S,H\right] \mid H\right] \quad \text{(Law of total expectation)} \\
&= \mathbb{E}\left[\mathbb{E}\left[\mu(A,O)\left(R+\gamma b_V(F')\right) - b_V(F) \mid S\right] \mid H\right].
\end{aligned}$$

In the last line, we use $(H) \perp (A,O,F') \mid S$. From (7b), we have

$$\mathbb{E}[b_V(F) \mid S] = \mathbb{E}\left[\mu(A,O)\left(R+\gamma b_V(F')\right) \mid S\right]. \tag{29}$$

Hence, $\mathcal{T}^{\mathcal{S}}(b_V)(S) = 0$. Then, from the overlap condition (7c),

$$\mathcal{T}^{\mathcal{S}}(b_V)(S^{\diamond}) = 0. \tag{30}$$

Finally, for any $b_V \in \mathcal{B}_V \cap \mathcal{Q}$, we have

$$\begin{aligned}
&(J(\pi^e) - \mathbb{E}_{f \sim \nu_{\mathcal{F}}}[b_V(f)]) \\
&= (1-\gamma)^{-1}\mathbb{E}_{(\tilde s)\sim d_{\pi^e}}\left[\mathbb{E}[\mu(A,O)R \mid S = \tilde s]\right] - \mathbb{E}_{f \sim \nu_{\mathcal{F}}}[b_V(f)] \\
&= (1-\gamma)^{-1}\mathbb{E}_{(\tilde s)\sim d_{\pi^e}}\left[\mathbb{E}\left[\mu(A,O)\{R+\gamma b_V(F')\} - b_V(F) \mid S = \tilde s\right]\right] \quad \text{(Use Lemma 9)} \\
&= 0. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\text{Use } \mathcal{T}^{\mathcal{S}}(S^{\diamond}) = 0.)
\end{aligned}$$

From the first line to the second line, we use

$$\begin{aligned}
J(\pi^e) &= (1-\gamma)^{-1}\int d_{\pi^e}(z,s)r(s,a)\pi^e(a \mid z,o)\mathrm{d}(z,s) \\
&= (1-\gamma)^{-1}\mathbb{E}_{(\tilde s)\sim d_{\pi^e}}\left[\mathbb{E}[\mu(A,O)R \mid S = \tilde s]\right].
\end{aligned}$$

L.2 Proof of Lemma 4

From (LM2), there exists w_1 such that $V^{\pi^e}(S) = w_1^\top \phi_{\mathcal{S}}(S)$. Then, from (LM1),

$$\mathbb{E}[w_2^\top \phi_{\mathcal{F}}(F) \mid S] = w_2^\top K_1 \phi_{\mathcal{S}}(S) = w_1^\top \phi_{\mathcal{S}}(S).$$

From (LM3), the above equation has a solution with respect to w_2. This concludes the proof.

L.3 Proof of Lemma 5

From (LM1), (LM4) and the statement of Lemma 4, there exists $w_4 \in \mathbb{R}^{d_{\mathcal{S}}}$ such that

$$\mathbb{E}[\mu(O,A)\{R + \gamma q(F')\} - q(F) \mid S] = w_4^\top \phi_{\mathcal{S}}(S).$$

for any $q(\cdot) = w^\top \phi_{\mathcal{F}}(\cdot) \in \mathcal{Q}$. Letting $b_F(\cdot) = \{w^\star\}^\top \phi_{\mathcal{F}}(\cdot)$, this is because

$$\begin{aligned}
&\mathbb{E}[\mu(O,A)\{R + \gamma q(F')\} - q(F) \mid S] && (31)\\
&= \mathbb{E}[\mu(O,A)\{\gamma q(F') - \gamma b_V(F')\} - q(F) + b_V(F) \mid S] && \text{(Statement of Lemma 4)}\\
&= \mathbb{E}[\mu(O,A)\{\gamma(w^\top - \{w^\star\}^\top)\mathbb{E}[\phi_{\mathcal{F}}(F') \mid S']\} - (w^\top - \{w^\star\}^\top)\mathbb{E}[\phi_{\mathcal{F}}(\mathcal{F}) \mid S] \mid S] &&\\
&= \mathbb{E}[\mu(O,A)\{\gamma(w^\top - \{w^\star\}^\top)K_1\phi_{\mathcal{S}}(S')\} - (w^\top - \{w^\star\}^\top)K_1\phi_{\mathcal{S}}(S) \mid S] && \text{((LM1))}\\
&= w_4^\top \phi_{\mathcal{S}}(S). && \text{((LM4))}
\end{aligned}$$

for some w_4.

Then, from (LM5), when $w_4^\top \mathbb{E}[\phi_{\mathcal{S}}(S) \mid H] = 0$, we have $w_4^\top \phi_{\mathcal{S}}(S) = 0$. This is because first $\mathbb{E}[w_4^\top \phi_{\mathcal{S}}(S) \mid H] = 0$ implies $\mathbb{E}[\{\mathbb{E}[w_4^\top \phi_{\mathcal{S}}(S) \mid H]\}^2] = 0$. Then, to make the ratio

$$\sup_{w \in \mathbb{R}^d} \frac{\mathbb{E}[\{w^\top \phi_{\mathcal{S}}(S)\}^2]}{\mathbb{E}[\{w^\top \mathbb{E}[\phi_{\mathcal{S}}(S) \mid H]\}^2]}$$

finite, we need $\mathbb{E}[\{w_4^\top \phi_{\mathcal{S}}(S)\}^2] = 0$. This implies $w_4^\top \phi_{\mathcal{S}}(S) = 0$.

Here, when we have

$$\begin{aligned}
0 &= \mathbb{E}[\mathbb{E}[\mu(O,A)\{R + \gamma q(F')\} - q(F) \mid S] \mid H] &&\\
&= \mathbb{E}[w_4^\top \phi_{\mathcal{S}}(S) \mid H], && \text{(Just plug-in)}
\end{aligned}$$

we get $w_4^\top \phi_{\mathcal{S}}(S) = 0$, i.e.,

$$\mathbb{E}[\mu(O,A)\{R + \gamma q(F')\} - q(F) \mid S] = 0.$$

L.4 Proof of Lemma 6

Letting

$$\tilde{w} = \mathbb{E}[\phi_{\mathcal{H}}(H)\{\gamma\phi_{\mathcal{F}}(F') - \phi_{\mathcal{F}}(F)\}]^{+}\mathbb{E}[\mu(O,A)R\phi_{\mathcal{H}}(H)],$$

we want to prove $\tilde{w}^\top \phi_{\mathcal{F}}(\cdot)$ is a learnable future-dependent value function. Then, by invoking Theorem 7, the statement is proved.

First step. Here, for $q \in \mathcal{Q}$, we have $(\mathcal{T}q)(H) = a^\top \phi_{\mathcal{H}}(H)$ for some vector $a \in \mathbb{R}^{d_{\mathcal{H}}}$. Here, $\mathbb{E}[(\mathcal{T}q)^2(H)] = 0$ is equivalent to $(\mathcal{T}q)(H) = 0$. Besides, the condition $\mathbb{E}[(\mathcal{T}q)^2(H)] = 0$ is equivalent to

$$a^\top \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{H}}^\top(H)]a = 0.$$

Thus, $\mathbb{E}[(\mathcal{T}q)(H)\phi_{\mathcal{H}}^\top(H)] = a^\top \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{H}}^\top(H)] = \mathbf{0}$, where $\mathbf{0} \in \mathbb{R}^{d_{\mathcal{H}}}$ is a vector consisting of 0, is a sufficient condition to satisfy $(\mathcal{T}q)(H) = \mathbf{0}$. Hence, if $q(\cdot) = w^\top \phi_{\mathcal{F}}(\cdot) \in \mathcal{Q}$ satisfies

$$\mathbb{E}[\phi_{\mathcal{H}}(H)\{\mu(O,A)\{R + \gamma q(F')\} - q(F)\}] = \mathbf{0}, \tag{32}$$

$q(\cdot)$ is a learnable bridge function. Note the above equation is equal to

$$\mathbb{E}[\phi_{\mathcal{H}}(H)\{\gamma\mu(O,A)\phi_{\mathcal{F}}(F') - \phi_{\mathcal{F}}(F)\}^\top]w = \mathbb{E}[\mu(O,A)R\phi_{\mathcal{H}}(H)].$$

Vice versa, i.e., any learnable future-dependent value function satisfies the above (32) is similarly proved.

Second step. Since there exists a linear learnable future-dependent value function in $\mathcal{Q}$ from Lemma 4, we have a solution to

$$\mathbb{E}[\phi_{\mathcal{H}}(H)\{\gamma\mu(O,A)\phi_{\mathcal{F}}(F') - \phi_{\mathcal{F}}(F)\}^\top]w = \mathbb{E}[\mu(O,A)R\phi_{\mathcal{H}}(H)]$$

with respect to w. We denote it by w. Thus, $\tilde{w}$ is also a solution since

$$B\tilde{w} = BB^+Bw = Bw = \mathbb{E}[\mu(O,A)R\phi_{\mathcal{H}}(H)]$$

where $B = \mathbb{E}[\phi_{\mathcal{H}}(H)\{\gamma\phi_{\mathcal{F}}(F') - \phi_{\mathcal{F}}(F)\}^\top]$. We use $B = BB^+B$. Note $\tilde{w}$ and w can be different.

L.5 Proof of Lemma 7

Refer to [USL+22, Chapter J].

M Proof of Section E

M.1 Proof of Theorem 8

Take $b_V^{[t]} \in \mathcal{B}_V^{[t]} \cap \mathcal{Q}_t$. Then, we have

$$\mathbb{E}[\mu(O,A)\{R + \gamma b_V^{[t+1]}(F')\} - b_V^{[t]}(F) \mid H] = 0.$$

Here, this implies

$$\mathbb{E}[\mathbb{E}[\mu(O,A)\{R + \gamma b_V^{[t+1]}(F')\} - b_V^{[t]}(F) \mid S] \mid H] = 0$$

using $H \perp (A, O, R, F') \mid S$.

Then, from the we have $\mathcal{T}_t^{\mathcal{S}}(b_V)(S) = 0$, i.e.,

$$\mathbb{E}[\mu(O,A)\{R + \gamma b_V^{[t+1]}(F')\} - b_V^{[t]}(F) \mid S] = 0.$$

using the assumption (b). Next, from the overlap assumption (c), we have $\{\mathcal{T}^{\mathcal{S},t}(b_V)\}(S_t^\diamond) = 0$ where $S_t^\diamond \sim d_t^{\pi^e}(\cdot)$.

Therefore, we have

$$\begin{aligned}
&J(\pi^e) - \mathbb{E}_{f\sim\nu_{\mathcal{F}}}[b_V^{[0]}(f)] \\
&= \left(\sum_{t=0}^{T-1} \mathbb{E}_{s\sim d_t^{\pi^e}}[\gamma^t \mathbb{E}[\mu(O,A)R \mid S = s]\right) + \left(\sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{s\sim d_{t+1}^{\pi^e}}[\mathbb{E}[b_V^{[t+1]}(F) \mid S = s]] - \mathbb{E}_{s\sim d_t^{\pi^e}}[\mathbb{E}[b_V^{[t]}(F) \mid S = s]]\right) && \text{(Telescoping sum)}
\end{aligned}$$

from telescoping sum. Besides, we have

$$\begin{aligned}
&\mathbb{E}_{s\sim d_{t+1}^{\pi^e}}[\mathbb{E}[b_V^{[t+1]}(F) \mid S = s]] \\
&= \mathbb{E}_{s\sim d_t^{\pi^e}}[\mathbb{E}[\mu(O,A)\mathbb{E}[b_V^{[t+1]}(F) \mid S'] \mid S = s]] \\
&= \mathbb{E}_{s\sim d_t^{\pi^e}}[\mathbb{E}[\mu(O,A)\mathbb{E}[b_V^{[t+1]}(F) \mid S', O, A] \mid S = s]] && (F \perp (O,A) \setminus S' \mid S') \\
&= \mathbb{E}_{s\sim d_t^{\pi^e}}[\mathbb{E}[\mu(O,A)b_V^{[t+1]}(F) \mid S = s]]. && \text{(Total law of expectation)}
\end{aligned}$$

Therefore,

$$\begin{aligned}
&J(\pi^e) - \mathbb{E}_{f\sim\nu_{\mathcal{F}}}[b_V^{[0]}(f)] \\
&= \left(\sum_{t=0}^{T-1} \mathbb{E}_{s\sim d_t^{\pi^e}}[\gamma^t \mathbb{E}[\mu(O,A)R \mid S = s]\right) + \left(\sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{s\sim d_t^{\pi^e}}[\mathbb{E}[\gamma\mu(O,A)b_V^{[t+1]}(F') - b_V^{[t]}(F) \mid S = s]]\right) \\
&= \sum_{t=0}^{T-1} \gamma^t \mathbb{E}_{s\sim d_t^{\pi^e}}[\mathbb{E}[\mu(O,A)\{R + \gamma b_V^{[t+1]}(F')\} - b_V^{[t]}(F) \mid S = s]] \\
&= \sum_{t=0}^{T-1} \mathbb{E}_{s\sim d_t^{\pi^e}}[\mathcal{T}_t^{\mathcal{S}} s] \\
&= 0. && \text{(Recall we derive } \{\mathcal{T}^{\mathcal{S},t}(b_V)\}(S_t^\diamond) = 0.\text{)}
\end{aligned}$$

M.2 Proof of Corollary 1

The proof consists of three steps. We use Theorem 8.

First step: verify existence of learnable future-dependent value functions (Show (8a)). From (LM1), we need to find a solution to

$$w_1^\top \mathbb{E}[\phi_{\mathcal{F}}(F) \mid S] = V_t^{\pi^e}(S)$$

with respect to a value w_1. Then, from (LM2f), there exists w_2 and K_1 such that

$$w_1^\top K_1 \phi_{\mathcal{S}}(S) = w_2^\top \phi_{\mathcal{S}}(S).$$

Thus, from (LM3), the above has a solution with respect to w_1.

Second step: verify invertibility condition (Show (8b)). Take a function $q^{[t]}(F)$ linear in $\phi_{\mathcal{F}}(F)$. From (LM1), (LM2) and (LM4), there exists $w_3 \in \mathbb{R}^{d_S}$ such that

$$\mathbb{E}[\mu(O,A)\{R + \gamma q^{[t+1]}(F') - q^{[t]}(F)\} \mid S] = w_3^\top \phi_{\mathcal{S}}(S).$$

This is proved as in (31). Then, $w_3^\top \mathbb{E}[\phi_{\mathcal{S}}(S) \mid H] = 0$ implies $w_3^\top \phi_{\mathcal{S}}(S) = 0$ from (LM5). Hence, when we have

$$\begin{aligned} 0 &= \mathbb{E}[\mathbb{E}[\mu(O,A)\{R + \gamma q^{[t+1]}(F') - q^{[t]}(F)\} \mid S] \mid H] \\ &= \mathbb{E}[w_3^\top \phi_{\mathcal{S}}(S) \mid H], \end{aligned}$$

this implies $w_3^\top \phi_{\mathcal{S}}(S) = 0$, i.e.,

$$\mathbb{E}[\mu(O,A)\{R + \gamma q^{[t+1]}(F') - q^{[t]}(F)\} \mid S] = 0.$$

Third step: show learnable future-dependent value functions are future-dependent value functions. Take a learnable future-dependent value function b_V. Then, from the condition, we have

$$\mathbb{E}[\mu(O,A)\{R + \gamma b_V^{[t+1]}(F')\} - b_V^{[t]}(F) \mid S] = 0.$$

We want to prove

$$\mathbb{E}[b_V^{[t]}(F) \mid S] = V_t^{\pi^e}(S).$$

We use induction. When $t = T-1$, this is clear. Next, supposing the statement is true at $t+1$, we prove it at a horizon t. Here, we have

$$\begin{aligned} \mathbb{E}[b_V^{[t]}(F) \mid S] &= \mathbb{E}[\mu(O,A)\{R + \gamma b_V^{[t+1]}(F')\} \mid S] \\ &= \mathbb{E}[\mu(O,A)\{R + \gamma \mathbb{E}[b_V^{[t+1]}(F') \mid S']\} \mid S] \\ &= \mathbb{E}[\mu(O,A)\{R + \gamma V_{t+1}^{\pi^e}(S')\} \mid S] = V_t^{\pi^e}(S). \end{aligned}$$

Thus, from induction, we have $\mathbb{E}[b_V^{[t]}(F) \mid S] = V_t^{\pi^e}(S)$ for any $t \in [T-1]$ for any learnable future-dependent value function $b_V(\cdot)$.

Fourth step: show the final formula. Recall we define

$$\theta_t = \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]^+ \mathbb{E}[\mu(O,A)\phi_{\mathcal{H}}(H)\{R + \gamma \phi_{\mathcal{F}}^\top(F')\theta_{t+1}\}]$$

We want to show $\theta_t^\top \phi_{\mathcal{F}}(\cdot)$ is a learnable future-dependent value function. Here, we need to say

$$\mathbb{E}[\mu(O,A)\{R + \gamma \theta_{t+1}^\top \phi_{\mathcal{F}}(F')\} - \theta_t^\top \phi_{\mathcal{F}}(F) \mid H] = 0.$$

This is satisfied if we have

$$\mathbb{E}\left(\phi_{\mathcal{H}}(H)\left(\mu(O,A)\{R + \gamma \phi_{\mathcal{F}}^\top(F')\theta_{t+1}\} - \phi_{\mathcal{F}}^\top(F)\theta_t\right)\right) = 0. \tag{33}$$

This is because $\mathbb{E}[\mu(O,A)\{R + \gamma \theta_{t+1}^\top \phi_{\mathcal{F}}(F')\} - \theta_t^\top \phi_{\mathcal{F}}(F) \mid H] = w_5^\top \phi_{\mathcal{H}}(H)$ for some vector w_5. Besides, $\mathbb{E}[w_5^\top \phi_{\mathcal{H}}(H)\phi_{\mathcal{H}}(H)] = 0$ implies $\mathbb{E}[w_5^\top \phi_{\mathcal{H}}(H)\phi_{\mathcal{H}}(H) w_5] = 0$, which results in

$$w_5^\top \phi_{\mathcal{H}}(H) = 0.$$

On top of that, since future-dependent value functions $\langle \tilde{\theta}_t, \phi_{\mathcal{F}}(F) \rangle$ exist from the first statement, we have a solution:

$$\mathbb{E}\left(\phi_{\mathcal{H}}(H)\left(\mu(O,A)\{R+\gamma\phi_{\mathcal{F}}^{\top}(F')\tilde{\theta}_{t+1}\} - \phi_{\mathcal{F}}^{\top}(F)\tilde{\theta}_t\right)\right) = 0.$$

with respect to $\tilde{\theta}_t$. In the following, we use this fact.

Now, we go back to the proof of the main statement, i.e., we prove (33). We use the induction. This is immediately proved when $t = T-1$. Here, suppose $\phi_{\mathcal{F}}^{\top}(\cdot)\theta_{t+1}$ is a learnable future-dependent value function at $t+1$. Then, we have

$$\begin{aligned}
&\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^{\top}]\theta_t \\
&= \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^{\top}]\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^{\top}]^{+}\mathbb{E}[\mu(O,A)\phi_{\mathcal{H}}(H)\{R+\gamma\phi_{\mathcal{F}}^{\top}(F')\theta_{t+1}\}] \quad \text{(Definition)} \\
&= \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^{\top}]\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^{\top}]^{+}\mathbb{E}[\mu(O,A)\phi_{\mathcal{H}}(H)\{R+\gamma V_{t+1}^{\pi^e}(S')\}].
\end{aligned}$$

In the last line, we use the inductive hypothesis and the third step. Recall we showed in the previous step $\mathbb{E}[b_V^{[t]}(F) \mid S] = V_t^{\pi^e}(S)$ for any learnable future-dependent value functions $b_V^{[t]}(\cdot)$. Then, we have

$$\begin{aligned}
&\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^{\top}]\theta_t \\
&= \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^{\top}]\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^{\top}]^{+}\mathbb{E}[\mu(O,A)\phi_{\mathcal{H}}(H)\{R+\gamma V_{t+1}^{\pi^e}(S')\}] \\
&= \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^{\top}]\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^{\top}]^{+}\mathbb{E}[\mu(O,A)\phi_{\mathcal{H}}(H)\{R+\gamma\phi_{\mathcal{F}}^{\top}(F')\tilde{\theta}_{t+1}\}] \\
&= \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^{\top}]\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^{\top}]^{+}\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^{\top}]\tilde{\theta}_t \\
&= \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^{\top}]\tilde{\theta}_t \qquad \text{(Property of Moore-Penrose Inverse)} \\
&= \mathbb{E}[\mu(O,A)\phi_{\mathcal{H}}(H)\{R+\gamma\phi_{\mathcal{F}}^{\top}(F')\tilde{\theta}_{t+1}\}] \qquad \text{(Definition of } \tilde{\theta}_t\text{)} \\
&= \mathbb{E}[\mu(O,A)\phi_{\mathcal{H}}(H)\{R+\gamma V_{t+1}^{\pi^e}(S')\}] \qquad \text{(We showed in the previous step)} \\
&= \mathbb{E}[\mu(O,A)\phi_{\mathcal{H}}(H)\{R+\gamma\phi_{\mathcal{F}}^{\top}(F')\theta_{t+1}\}]. \qquad \text{(From the induction)}
\end{aligned}$$

Hence, $\phi_{\mathcal{F}}^{\top}(\cdot)\theta_t$ is a learnable future-dependent value function at t.

N Proof of Section F

N.1 Proof of Theorem 9

We take a value bridge function $b_D^{[t]} \in \mathcal{Q}_t \cap \mathcal{B}_D^{[t]}$ for any $[T]$. Then, we define

$$l_D^{[t]}(\cdot) = \mathbb{E}[b_D^{[t]}(F) \mid S = \cdot].$$

In this section, $\mathbb{E}[\cdot\,; A_{0:T-1} \sim \pi^e]$ means taking expectation when we execute a policy π_e from $t=0$ to $T-1$. Note $\mathbb{E}[\cdot\,; A_{0:T-1} \sim \pi^b]$ is just $\mathbb{E}[\cdot]$. Here, we have

$$\begin{aligned}
&\mathrm{Pr}_{\pi^e}(o_0, a_0, \cdots, a_{T-1}) - \mathbb{E}_{f\sim\nu_{\mathcal{F}}}[b_D^{[0]}(f)] \\
&= \sum_{t=0}^{T-1}\mathbb{E}\left[\left\{\prod_{k=0}^{t-1}\mathrm{I}(O_k = o_k, A_k = a_k)\right\}\left\{\mathrm{I}(O_t = o_t, A_t = a_t) l_D^{[t+1]}(S_{t+1}) - l_D^{[t]}(S_t)\right\}; A_{0:T-1} \sim \pi^e\right] \\
&= \sum_{t=0}^{T-1}\mathbb{E}[\Bigg\{\prod_{k=0}^{t-1}\mathrm{I}(O_k = o_k, A_k = a_k)\Bigg\}\{\mathbb{E}[\mathrm{I}(O_t = o_t, A_t = a_t) l_D^{[t+1]}(S_{t+1}) \mid S_t, (O_0, A_0, \cdots, A_{t-1}) = (o_0, a_0, \cdots, a_{t-1})] \\
&- l_D^{[t]}(S_t)\}; A_{0:T-1} \sim \pi^e] \\
&= \sum_{t=0}^{T-1}\mathbb{E}\left[\left\{\prod_{k=0}^{t-1}\mathrm{I}(O_k = o_k, A_k = a_k)\right\}\left\{\mathbb{E}[\mathrm{I}(O_t = o_t, A_t = a_t) l_D^{[t+1]}(S_{t+1}) \mid S_t] - l_D^{[t]}(S_t)\right\}; A_{0:T-1} \sim \pi^e\right].
\end{aligned}$$

In the first line, we use a telescoping sum trick noting

$$\mathrm{Pr}_{\pi^e}(o_0, a_0, \cdots, a_{T-1}) = \mathbb{E}\left[\prod_{k=0}^{T-1}\mathrm{I}(O_k = o_k, A_k = a_k); A_{0:T-1} \sim \pi^e\right].$$

From the second line to the third line, we use $S_{t+1}, O_t, A_t \perp O_1, \cdots, A_t \mid S_t$.

Here, we have

$$\mathbb{E}[\mathrm{I}(O_t = o_t, A_t = a_t) l_D^{[t+1]}(S_{t+1}) \mid S_t; A_t \sim \pi^e] = \tilde{l}_D^{[t]}(S_t)$$

where $\tilde{l}_D^{[t+1]}(\cdot) = \mathbb{E}[\mathrm{I}(O = o_t, A = a_t)\mu(O, A) l_D^{[t+1]}(S') \mid S = \cdot; A_t \sim \pi^b]$ using importance sampling. Hence, the following holds:

$$\begin{aligned}
&\mathrm{Pr}_{\pi^e}(o_0, a_0, \cdots, a_{T-1}) - \mathbb{E}_{f \sim \nu_{\mathcal{F}}}[b_D^{[0]}(f)] \\
&= \sum_{t=0}^{T-1} \mathbb{E}\left[\left\{\prod_{k=0}^{t-1} \mathrm{I}(O_k = o_k, A_k = a_k)\right\} \left\{\tilde{l}_D^{[t+1]}(S_t) - l_D^{[t]}(S_t)\right\}; A_{0:T-1} \sim \pi^e\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}\left[\left\{\prod_{k=0}^{t-1} \mathrm{I}(O_k = o_k, A_k = a_k)(\mathcal{T}_t^{\mathcal{S}} b_D)(S_t)\right\}; A_{0:T-1} \sim \pi^e\right] \qquad (34)\\
&= \sum_{t=0}^{T-1} \mathbb{E}\left[\left\{\prod_{k=0}^{t-1} \mathrm{I}(O_k = o_k, A_k = a_k)\mathbb{E}[(\mathcal{T}_t^{\mathcal{S}} b_D)(S_t) \mid (O_1, A_1, \cdots, A_K) = (o_1, a_1, \cdots, o_K)]\right\}; A_{0:T-1} \sim \pi^e\right] \\
&\qquad (35)
\end{aligned}$$

From the second line to the third line, we use

$$\begin{aligned}
&\mathbb{E}[\mathrm{I}(O = o_t, A = a_t)\mu(O, A) l_D^{[t+1]}(S') \mid S] \\
&= \mathbb{E}[\mathrm{I}(O = o_t, A = a_t)\mu(O, A)\mathbb{E}[b_D^{[t+1]}(F') \mid S'] \mid S] && \text{(Definition)} \\
&= \mathbb{E}[\mathrm{I}(O = o_t, A = a_t)\mu(O, A)\mathbb{E}[b_D^{[t+1]}(F') \mid S', O = o_t, A = a_t] \mid S] \\
& && \text{(Low of total expectation)} \\
&= \mathbb{E}[\mathrm{I}(O = o_t, A = a_t)\mu(O, A) b_D^{[t+1]}(F') \mid S]. && (F' \perp O, A \mid S')
\end{aligned}$$

Besides, we know for any learnable bridge function b_D, we have

$$\mathbb{E}[\mathrm{I}(O = o_t, A = a_t)\mu(O, A) b_D^{[t+1]}(F') - b_D^{[t]}(F) \mid H] = 0.$$

Then, since

$$\mathbb{E}[\mathbb{E}[\mathrm{I}(O = o_t, A = a_t)\mu(O, A) b_D^{[t+1]}(F') - b_D^{[t]}(F) \mid S] \mid H] = 0$$

from the invertibility condition (b), we have

$$\mathbb{E}[\mathrm{I}(O = o_t, A = a_t)\mu(O, A) b_D^{[t+1]}(F') - b_D^{[t]}(F) \mid S] = 0.$$

From the overlap (c), this implies

$$(\mathcal{T}_t^{\mathcal{S}} b_D)(\grave{S}_t) = 0.$$

Finally, from (35), we can conclude $\mathrm{Pr}_{\pi^e}(o_0, a_0, \cdots, a_{T-1}) - \mathbb{E}_{f \sim \nu_{\mathcal{F}}}[b_D^{[0]}(f)] = 0$.

N.2 Proof of Theorem 10

This is proved as in Theorem 9 noting

$$\mathrm{Pr}_{\pi^e}(o_0, a_0, \cdots, a_{T-1}, \mathbf{O}_T) = \mathbb{E}\left[\phi_{\mathcal{O}}(O_T) \prod_{k=0}^{T-1} \mathrm{I}(O_k^{\diamond} = o_k, A_k^{\diamond} = a_k)\right].$$

N.3 Proof of Corollary 2

This is proved following Corollary 3.

N.4 Proof of Corollary 3

The proof consists of four steps. The proof largely follows the proof of Corollary 1.

First step: verify the existence of bridge functions (9a). We show there exists a bridge function linear in $\phi_{\mathcal{F}}(F)$. We need to find a solution to $w_1^\top \mathbb{E}[\phi_{\mathcal{F}}(F) \mid S] = V^{\pi^e}_{D,[t]}(S)$ with respect to w_1. Then, from (LM2D), there exists w_2 and K_1 such that

$$w_1^\top K_1 \phi_{\mathcal{S}}(S) = w_2^\top \phi_{\mathcal{S}}(S).$$

Thus, from (LM3), the above has a solution with respect to w_1.

Second step: verify invertibility conditions (9b). We take $\mathcal{Q}_t$ to be a linear model in $\phi_{\mathcal{F}}(F)$. For $q^{[t]}_D \in \mathcal{Q}_t$, from (LM1), (LM2), (LM4), there exists $w_3 \in \mathbb{R}^{d_{\mathcal{S}}}$ such that

$$\mathbb{E}[q^{[t]}_D(F) - \mathbb{I}(O = o_t, A = a_t)\mu(O,A) q^{[t+1]}_D(F') \mid S] = w_3^\top \phi_{\mathcal{S}}(S).$$

Then, $w_3^\top \mathbb{E}[\phi_{\mathcal{S}}(S) \mid H] = 0$ implies $w_3^\top \phi_{\mathcal{S}}(S) = 0$ from (LM5). Hence,

$$\begin{aligned} 0 &= \mathbb{E}[q^{[t]}_D(F) - \mathbb{I}(O = o_t, A = a_t)\mu(O,A) q^{[t+1]}_D(F') \mid H] \\ &= \mathbb{E}[\mathbb{E}[q^{[t]}_D(F) - \mathbb{I}(O = o_t, A = a_t)\mu(O,A) q^{[t+1]}_D(F') \mid S] \mid H] \end{aligned}$$

implies $\mathbb{E}[q^{[t]}_D(F) - \mathbb{I}(O = o_t, A = a_t)\mu(O,A) q^{[t+1]}_D(F') \mid S] = 0$. Thus, the invertibility is concluded.

Third step: show learnable value bridge functions are value bridge functions. From the previous discussion, learnable value bridge functions in $\mathcal{Q}$ need to satisfy

$$0 = \mathbb{E}[b^{[t]}_D(F) - \mathbb{I}(O = o_t, A = a_t)\mu(O,A) b^{[t+1]}_D(F') \mid S].$$

Here, we want to prove

$$\mathbb{E}[b^{[t]}_D(F) \mid S] = V^{\pi^e}_{D,[t]}(S).$$

We use induction. When $t = T-1$, this is clear. Next, supposing the statement is true at $t+1$, we prove it at horizon t. Here, we have

$$\begin{aligned} \mathbb{E}[b^{[t]}_D(F) \mid S] &= \mathbb{E}[\mathbb{I}(O = o_t, A = a_t)\mu(O,A) b^{[t+1]}_D(F') \mid S] \\ &= \mathbb{E}[\mathbb{I}(O = o_t, A = a_t)\mu(O,A) \mathbb{E}[b^{[t+1]}_D(F') \mid S'] \mid S] \\ &= \mathbb{E}[\mathbb{I}(O = o_t, A = a_t)\mu(O,A) V^{\pi^e}_{D,[t+1]}(S') \mid S] \\ &= V^{\pi^e}_{D,[t]}(S). \end{aligned}$$

Fourth step: show the final formula. We recursively define

$$\theta_t^\top = \theta_{t+1}^\top D_t B^+.$$

starting from $\tilde{w}_T^\top B = \mathbb{E}[\phi_{\mathcal{H}}(H)]$. We want to show $\theta_t^\top \phi_{\mathcal{F}}(\cdot)$ is a learnable value bridge function. Here, we want to say

$$\mathbb{E}[\theta_t^\top \phi_{\mathcal{F}}(F) - \mathbb{I}(O = o_t, A = a_t)\mu(O,A)\theta_{t+1}^\top \phi_{\mathcal{F}}(F') \mid H] = 0.$$

This is satisfied if we have

$$\mathbb{E}\left[\left(\theta_t^\top \phi_{\mathcal{F}}(F) - \mathbb{I}(O = o_t, A = a_t)\mu(O,A)\theta_{t+1}^\top \phi_{\mathcal{F}}(F')\right)\phi_{\mathcal{H}}(H)\right] = 0$$

Here, refer to the fourth step in the proof of Corollary 1. Besides, as we already show linear value bridge functions exist $\langle \tilde{\theta}_t, \phi_{\mathcal{F}}(\cdot)\rangle$, we have a solution to:

$$\mathbb{E}\left[\left(\tilde{\theta}_t^\top \phi_{\mathcal{F}}(F) - \mathbb{I}(O = o_t, A = a_t)\mu(O,A)\tilde{\theta}_{t+1}^\top \phi_{\mathcal{F}}(F')\right)\phi_{\mathcal{H}}(H)\right] = 0$$

with respect to $\tilde{\theta}_t$.

Hereafter, we use the induction. This is immediately proved when $t = T-1$. Here, suppose $\phi_{\mathcal{F}}^\top(F)\theta_{t+1}$ is a learnable value bridge function at $t+1$. Then, we have

$$\begin{aligned} &\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]\theta_t \\ &= \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]^+ D_t^\top \theta_{t+1} \\ &= \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]^+ \mathbb{E}[\mathrm{I}(O = O_t, A = a_t)\mu(O,A)\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}^\top(F')\theta_{t+1}] \\ &= \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]^+ \mathbb{E}[\mathrm{I}(O = O_t, A = a_t)\mu(O,A)\phi_{\mathcal{H}}(H)V^{\pi^e}_{D,[t+1]}(S')]. \end{aligned}$$

In the last line, we use inductive hypothesis. We showed in the previous step $\mathbb{E}[b_D^{[t]}(F) \mid S] = V_{D,[t]}^{\pi^e}(S)$ for any learnable value bridge function $b_D^{[t]}(F)$. Then, we have

$$\begin{aligned}
&\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]^+\mathbb{E}[\mathrm{I}(O=O_t, A=a_t)\mu(O,A)\phi_{\mathcal{H}}(H)V_{D,[t+1]}^{\pi^e}(S')] \\
&= \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]^+\mathbb{E}[\mathrm{I}(O=O_t, A=a_t)\mu(O,A)\phi_{\mathcal{H}}(H)\{\phi_{\mathcal{F}}^\top(F')\tilde{\theta}_{t+1}\}] \\
&= \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]^+\mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]\tilde{\theta}_t \\
&= \mathbb{E}[\phi_{\mathcal{H}}(H)\phi_{\mathcal{F}}(F)^\top]\tilde{\theta}_t \qquad \text{(Property of Moore-Penrose Inverse)} \\
&= \mathbb{E}[\mathrm{I}(O=O_t, A=a_t)\mu(O,A)\phi_{\mathcal{H}}(H)V_{D,[t+1]}^{\pi^e}(S')] \qquad \text{(We show in the previous step)} \\
&= \mathbb{E}[\mathrm{I}(O=O_t, A=a_t)\mu(O,A)\phi_{\mathcal{H}}(H)\{\phi_{\mathcal{F}}^\top(F')\theta_{t+1}\}]. \qquad \text{(From the induction)}
\end{aligned}$$

So far, we prove $\theta_t^\top\phi_{\mathcal{F}}(\cdot)$ is a learnable value bridge function for t. Finally, by consdiering a time-step $t=0$, we can prove the target estimand is

$$\mathbb{E}[\tilde{w}_0^\top\phi_{\mathcal{F}}(F)] = \mathbb{E}[\phi_{\mathcal{H}}(H)]^\top B^+\left\{\prod_{t=T-1}^{0} D_t B^+\right\}C.$$

O Auxiliary Lemmas

Lemma 9. *Take $g \in [\mathcal{S} \to \mathbb{R}]$. Then, we have*

$$0 = (1-\gamma)^{-1}\mathbb{E}_{s\sim d_{\pi^e}}\left[\mathbb{E}[\mu(O,A)\gamma g(S') - g(S) \mid S=s]\right] + \mathbb{E}_{s\sim\nu_S}[gs].$$

Proof. Let $d_{\pi^e}(z,s) = d_0^{\pi^e}(\cdot)$. Then, we have

$$\begin{aligned}
&\int g(z,s)d_{\pi^e}(z,s)d(z,s) \\
&= (1-\gamma)\int g(z,s)\sum_{t=0}^{\infty}\gamma^t d_t^{\pi^e}(z,s)d(z,s) \\
&= \underbrace{(1-\gamma)\int g(z,s)d_0^{\pi^e}(z,s)d(z,s)}_{(a)} + \underbrace{\gamma(1-\gamma)\int g(z',s')\sum_{t=1}^{\infty}\gamma^{t-1}d_t^{\pi^e}(z',s')d(z',s')}_{(b)}.
\end{aligned}$$

We analyze the first term (a) and the second term (b). The first term (a) is $\mathbb{E}_{s\sim\nu_S}[gs]$. In the following, we analyze the second term.

Here, note

$$d_t^{\pi^e}(z',s') = \int \mathbb{T}(s' \mid s,a)\mathbb{O}(o \mid s)\pi^e(a \mid z,o)\delta^\dagger(z'=z)d_{t-1}^{\pi^e}(z,s)d(z,s).$$

where $\delta^\dagger(z'=z) = \delta(f_{-1}(z') = f_{+1}(z))$. Here, f_{-1} is a transformation removing the most recent tuple (o,a,r) and f_{+1} is a transformation removing the oldest tuple (o,a,r). Hence,

$$\begin{aligned}
&(1-\gamma)\sum_{t=1}^{\infty}\gamma^{t-1}d_t^{\pi^e}(z',s') \\
&= (1-\gamma)\int\sum_{t=1}^{\infty}\gamma^{t-1}\mathbb{T}(s' \mid s,a)\mathbb{O}(o \mid s)\pi^e(a \mid z,o)\delta^\dagger(z'=z)d_{t-1}^{\pi^e}(z,s)d(z,s) \\
&= (1-\gamma)\int\sum_{k=0}^{\infty}\gamma^{k}\mathbb{T}(s' \mid s,a)\mathbb{O}(o \mid s)\mu(z,o,a)\pi^b(a \mid z,o)\delta^\dagger(z'=z)d_k^{\pi^e}(z,s)d(z,s) \\
&= \int\mathbb{T}(s' \mid s,a)\mathbb{O}(o \mid s)\mu(z,o,a)\pi^b(a \mid z,o)\delta^\dagger(z'=z)d^{\pi^e}(z,s)d(z,s).
\end{aligned}$$

Therefore, the term (b) is

$$
\begin{aligned}
&\gamma(1-\gamma)\int g(z',s')\sum_{t=1}^{\infty}\gamma^{t-1}d_t^{\pi^e}(z',s')d(z',s') \\
&=\gamma\int g(z',s')\mathbb{T}(s'\mid s,a)\mathbb{O}(o\mid s)\mu(z,o,a)\pi^b(a\mid z,o)\delta^{\dagger}(z'=z)d^{\pi^e}(z,s)d(z,z',s,s') \\
&=\gamma\mathbb{E}_{s\sim d_{\pi^e}}\left[\mathbb{E}[\mu(O,A)\gamma g(S')\mid S=s]\right].
\end{aligned}
$$

□