Off-Policy Evaluation for Missingness-Aware Policies in MDPs with Rewards Missing Not at Random

Anonymous Authors[1]

Abstract

In offline Reinforcement Learning, immediate rewards in logged batch data are often unobserved due to sparse or irregular record-keeping, or censored beyond certain reward values. This issue arises in practical settings, including health care and marketing. We investigate off-policy evaluation (OPE) in finite-horizon Markov decision processes when rewards are missing not at random (MNAR), which breaks ignorability and induces selection bias even after conditioning on states and actions. To address this, we formalize a reward-dependent propensity model and use future states as shadow variables to identify the full-data conditional mean reward. We further introduce a bridge function that recovers the conditional mean reward without explicitly modeling the MNAR mechanism, and estimate it via a min-max procedure to avoid double sampling. Building upon these identification results, we propose an Fitted-Q-Evaluation-style estimator that propagates the recovered rewards while allowing target policies to depend on past missingness indicators. Finally, we establish consistency and finite-sample error bounds for our OPE estimator, and show through simulations the strong performance of our method compared to existing benchmarks.

1. Introduction

Reinforcement Learning (RL) has achieved remarkable successes in sequential decision-making domains ranging from robotics to healthcare, and most recently in large language models and AI agents (Ouyang et al., 2022; Rafailov et al., 2023; Achiam et al., 2023). However, learning optimal policies often requires vast amounts of interaction with the environment, which can be costly, risky, and even unethical in high-stakes real world applications such as healthcare and education (Tsiatis et al., 2019; Murphy, 2003; Mandel et al., 2014). In these applications, we often have interaction data collected according to some behavior policy, e.g., standard care or usual strategy. Estimating the value of a target policy using historical datasets collected under a different behavior policy, a problem known as Off-Policy Evaluation (OPE), is essential in offline RL (Levine et al., 2020; Uehara et al., 2022b; Voloshin et al., 2021; Wang et al., 2024). Accurate OPE is critical for deploying safe and effective policies without the need for dangerous online exploration.

However, in many practical scenarios, such as medical treatment or digital advertising, the data is plagued by unobserved factors or missing values (Little & Rubin, 2019; Kallus & Zhou, 2020). A particularly pervasive challenge arises when rewards are missing not at random (MNAR), under which the probability of observing a reward depends on the latent value of the reward itself. For instance, in health care, patient records are often sparse and irregular; a patient may stop visiting the doctor when they feel recovered (a high unobserved reward) or, conversely, disengage to seek outside emergency care when their condition deteriorates severely (a low unobserved reward). Similarly, in multi-touch digital marketing, attribution is frequently disrupted by privacy limitations. While non-conversions (zero rewards) are trivially observable, high-value purchases often involve cross-device journeys, such as a user clicking on mobile but converting on a desktop, or trigger manual review flows that break attribution links. Consequently, high-value conversions go missing while low-value or null outcomes remain fully observed, creating a dataset that systematically biases the learning process against the most desirable outcomes. Similar MNAR feedback has been systematically studied in recommender systems, where popularity and exposure biases make logged interactions MNAR and consequently distort offline evaluations (Yang et al., 2018).

Standard OPE methods, such as Fitted Q-evaluation (FQE) and Importance Sampling (IS), rely on fully observed trajectories. Failing to account for the missingness could lead to biased OPE and hence sub-optimal decision-making. In the OPE literature, scheme of missingness has been studied in various aspects. Partially Observable Markov Decision

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Processes (POMDPs) (Kaelbling et al., 1998; Jaakkola et al., 1995) consider the case where the state is partially observed, which can be viewed as a special case of missingness. However, most existing POMDP methods assume that certain state variables are totally unobserved. In general, off-policy value in POMDPs are often unidentifiable without strong assumptions (Tennenholtz et al., 2020; Bennett et al., 2021; Shi et al., 2022; Uehara et al., 2022a). Some works define the missingness of certain status (e.g., hitting wall in Gridworld environment (Sutton & Barto, 2018)) in the reward function by assigning it to be some large negative values to discourage certain actions, which lead to certain states with voided rewards (Ng et al., 1999; Devlin & Kudenko, 2012). However, this approach may not accurately reflect the true reward structure and can introduce bias. Chu et al. (2023); Park et al. (2025) study the OPE problem with truncated trajectories, where they treat missingness as certain constraints. However, by penalizing on the missingness, these methods may shift the policy evaluation away from the true potential reward without missingness. Wang et al. (2025) propose an inverse probability weighting method for OPE with nonignorable truncation, but their method relies on an extra shadow variable, or requires expert knowledge to select such a variable from observed states.

In this paper, we study the OPE problem in MDPs with MNAR rewards to estimate values of target policies accounting for the past missingness. MNAR rewards break standard ignorability assumptions as the reward-dependent missingness induces selection bias and confounds state-action returns. The challenge is to recover the value of a target policy when the observed trajectories systematically underreport high or low rewards and when the missingness itself can depend on the past action and state, all without online data to re-collect or intervene.

We address these issues by formalizing the reward MNAR mechanism via a reward-dependent propensity score model and leveraging future states as shadow variables. Under mild completeness conditions, the shadow variables allow us to identify the full-data conditional mean reward even when the reward is MNAR. In addition, we introduce a bridge function $b_t(S_t, A_t, S_{t+1})$ satisfying $\mathbb{E}\{b_t(S_t, A_t, S_{t+1}) \mid R_t, S_t, A_t\} = R_t$, enabling recovery of the conditional mean reward without explicitly estimating the MNAR mechanism. This avoids the variance blow-up in inverse propensity weighting. We propose the min-max optimization to estimate the bridge function and the value function, which avoids the double sampling issue.

Building on these identification results, we further develop an FQE-style estimator that integrates the bridge function and allows target policies to depend on the previous missingness indicators. The procedure propagates the recovered rewards through the Bellman recursion of the target policy, yielding stable value estimates. We further establish the consistency and finite-sample error bounds of the proposed estimator in nonparametric settings. Extensive experiments on simulated data demonstrate the effectiveness of our method compared to existing benchmarks.

2. Related Work

OPE. Off-policy evaluation has been extensively studied in the RL literature. Classical methods include IS and its variants (Liu et al., 2018), FQE (Le et al., 2019), and doubly robust estimators that combine both (Kallus & Uehara, 2020). Recent advances in offline RL have developed pessimistic approaches (Xie et al., 2021; Rashidinejad et al., 2021; Shi et al., 2023; Zhan et al., 2022) that achieve near-optimal sample complexity. For comprehensive reviews, see Uehara et al. (2022b) and Levine et al. (2020). OPE in POMDPs has received growing attention, with works addressing latent confounding (Bennett et al., 2021; Kallus & Zhou, 2020), partial observability (Tennenholtz et al., 2020; Shi et al., 2022; Miao et al., 2022), and future-dependent estimation (Uehara et al., 2022a). However, none of these works directly address the MNAR reward setting.

Missing Data. Missing data problems have been extensively studied in statistics (Little & Rubin, 2019; Enders, 2022). Under missing at random (MAR) assumptions, inverse probability weighting and doubly robust methods are well-established. For MNAR, identification typically requires additional structure such as instrumental variables (Sun & Tchetgen Tchetgen, 2018), shadow variables (Zhao et al., 2015; Miao & Tchetgen Tchetgen, 2016; Miao & Tchetgen, 2018), or graphical constraints (Mohan & Pearl, 2021). Proximal causal inference (Tchetgen Tchetgen et al., 2020; Bennett & Kallus, 2021; Cui et al., 2024) has emerged as a powerful framework for handling unmeasured confounding using proxy variables. Our work extends these ideas to the OPE setting with MNAR rewards, leveraging future states as shadow variables for identification.

3. Preliminaries

We consider an episodic Markov Decision Process (MDP) $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \mathcal{P}, r, T\}$, where $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, respectively. The horizon length T is finite, and we assume the terminal state S_{T+1} is observed. The transition kernels $\mathcal{P} = \{P_t\}_{t=1}^T$ govern the state dynamics, where $P_t : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ maps state-action pairs to distributions over next states. The reward functions $r = \{r_t\}_{t=1}^T$ are defined as conditional expectations given the next state: $r_t(s, a, s') = \mathbb{E}[R_t \mid S_t = s, A_t = a, S_{t+1} = s']$ for any $(s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$. We assume bounded rewards $R_t \in \mathcal{R} \subseteq [-1, 1]$.

We introduce an observation indicator $O_t \in \{0,1\}$, where $O_t = 1$ indicates that the reward R_t is observed at time t, and $O_t = 0$ otherwise. Importantly, we allow the rewards to be missing not at random (MNAR), that is, even after conditioning on current states and actions, the missingness probability may depend on the possibly unobserved reward itself. We formalize this through the propensity score $e_t(s,a,r) = P(O_t = 1 \mid S_t = s, A_t = a, R_t = r)$ for $t = 1,\dots,T$.

Assumption 3.1 (No Future Dependence)**.** For all $t = 1,\dots,T-1$,

$$O_t \perp (S_{t+1:T}, R_{t+1:T}) \mid S_t, A_t, R_t.$$

This assumption states that the current missingness indicator O_t, given the current state, action, and reward, is independent of all future states and rewards. For example, in healthcare, whether a patient's health outcome is recorded typically depends on their current condition, not on future events that have not yet occurred.

Our goal is to evaluate the performance of a target policy $\pi = \{\pi_t\}_{t=1}^T$. We allow π_t to depend on the previous reward missingness, which is practically relevant when decisions adapt based on whether prior outcomes were observed. Formally, $\pi_t : \mathcal{S} \times \{0,1\} \to \Delta(\mathcal{A})$ with $\pi_t(a \mid s, o_-) = P(A_t = a \mid S_t = s, O_{t-1} = o_-)$. To accommodate this dependence, we define an augmented state $\widetilde{S}_t = (S_t, O_{t-1}) \in \widetilde{\mathcal{S}} = \mathcal{S} \times \{0,1\}$, so that the target policy can be written as $\pi_t(a \mid \widetilde{S}_t)$. The augmented process must satisfy the Markov property below.

Assumption 3.2 (Markov Property for Augmented Process)**.** The augmented process $\{(\widetilde{S}_t, A_t)\}_{t=1}^T$ with $\widetilde{S}_t = (S_t, O_{t-1})$ is an MDP

$$P(\widetilde{S}_{t+1} \mid \widetilde{S}_{1:t}, A_{1:t}) = P(\widetilde{S}_{t+1} \mid \widetilde{S}_t, A_t), \quad t = 1,\dots,T.$$

This assumption ensures that augmenting the state with the previous missingness indicator preserves the Markov property, and allows the value function recursion to hold with augmented state. The augmented transition kernel is $P(\widetilde{S}_{t+1} \mid \widetilde{S}_t, A_t) = P((S_{t+1}, O_t) \mid S_t, A_t)$. We set $O_0 = 0$ and let $S_1 \sim \rho_1$ denote the initial state distribution. The Q-function and value function satisfy the Bellman equation

$$\begin{aligned} Q_t^\pi(s,a) =& \mathbb{E}\big[r_t(s,a,S_{t+1}) + V_{t+1}^\pi(S_{t+1}, O_t) \mid S_t = s, \\ & A_t = a\big], \\ V_t^\pi(s,o_-) =& \sum_a \pi_t(a \mid s, o_-) Q_t^\pi(s,a), \quad V_{T+1}^\pi \equiv 0. \end{aligned} \tag{1}$$

The policy value is defined as $V(\pi) \equiv \mathbb{E}_{\widetilde{S}_1 \sim \tilde{\rho}_1}\big[V_1^\pi(\widetilde{S}_1)\big] = E_{S_1 \sim \rho_1}\big[V_1^\pi(S_1, 0)\big]$.

In OPE, data are collected under a behavior policy $\pi^b = \{\pi_t^b\}_{t=1}^T$, where $\pi_t^b : \mathcal{S} \to \Delta(\mathcal{A})$ does not depend on the missingness indicators. The observed dataset $\mathcal{D}$ consists of n i.i.d. trajectories $\tau_i = \{S_{t,i}, A_{t,i}, O_{t,i}, R_{t,i}^{\text{obs}}, S_{t+1,i}\}_{t=1}^T$ for $i = 1,\dots,n$, where $R_{t,i}^{\text{obs}} = O_{t,i} \cdot R_{t,i}$ denotes the observed reward (zero when missing). See Figure 1 for a directed acyclic graph (DAG) illustrating the data-generating process.

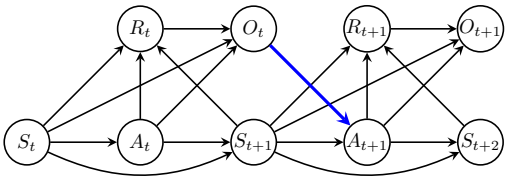

Figure 1. DAG for Target Policy and Behavior Policy. Black arrows represent the standard MDP dynamics and the MNAR reward mechanism. The blue arrow $O_t \to A_{t+1}$ indicates that the target policy may depend on the previous observation indicator O_t, whereas the behavior policy does not.

To deal with distribution shift in OPE, concentrability coefficients are often introduced (Munos, 2003; 2007; Chen & Jiang, 2019; Le et al., 2019; Duan et al., 2021). We define concentrability coefficient κ_t at time t in the following assumption.

Assumption 3.3 (Concentrability)**.** Given target policy π and behavior policy π^b, for each $t = 1,\dots,T$, assume there exist finite constants $\{\kappa_t\}_{t=1}^T$ such that

$$\Big\| \frac{\pi_t(a \mid s, o_-)}{\pi_t^b(a \mid s)} \Big\|_\infty \le \kappa_t.$$

Equivalently, for all (s, o_-, a) with $\pi_t^b(a \mid s) > 0$, $\pi_t(a \mid s, o_-) \le \kappa_t \, \pi_t^b(a \mid s)$.

Assumption 3.4. Assume there exist constants $a > 0$ and $\alpha_t \ge \alpha > 1$ such that for all $t \in \{1,\dots,T\}$,

$$\kappa_t \le 1 + \frac{a}{t^{\alpha_t}}.$$

Assumption 3.4 controls the cumulative growth of the concentrability coefficients so that the action mismatch does not compound over time.

Corollary 3.5 (Bounded cumulative concentrability)**.** *Under Assumptions 3.3 and 3.4, for $t = 1,\dots,T$,*

$$\Big(\prod_{j=1}^t \kappa_j\Big)^{1/2} \le \exp\Big(\frac{1}{2}\sum_{j=1}^t \frac{a}{j^{\alpha_j}}\Big) \le \exp\big(\frac{a}{2}\zeta(\alpha)\big) := K,$$

where $\zeta(\cdot)$ is the Riemann zeta function.

4. Identification

In this section, we establish identification results for the policy value, and formalize the conditions required for our approach.

To address the challenges posed by MNAR data in causal inference, Miao et al. (2015); Miao & Tchetgen Tchetgen (2016) propose identification methods with the help of auxiliary variables called *shadow variables*. Inspired by this line of research, we establish a nonparametric value-based approach for policy value identification. Our key insight is to adopt the next state S_{t+1} as the shadow variable, which serves as a proxy that helps recover information about the unobserved rewards. The shadow variable must satisfy two conditions that govern its relationship with the reward and missingness indicator.

Assumption 4.1 (Exclusion Restriction)**.** Suppose for all $t = 1, \dots, T$, S_{t+1} satisfies

$$S_{t+1} \perp O_t \mid R_t, S_t, A_t.$$

Assumption 4.2 (Relevance Condition)**.** Suppose for all $t = 1, \dots, T$, S_{t+1} satisfies

$$S_{t+1} \not\perp R_t \mid S_t, A_t, O_t = 1.$$

The two assumptions above are basic conditions for S_{t+1} to be a valid shadow variable at time t. Assumption 4.1 shows that conditional on the current state-action pair and the (possibly unobserved) reward, S_{t+1} provides no additional information about whether the reward is observed. Assumption 4.2 ensures that on the observed subset, S_{t+1} remains informative about R_t beyond what is already captured by (S_t, A_t). This condition guarantees that the shadow variable carries useful information about the reward. In the causal graph in Figure 1, Assumption 4.1 is consistent with d-separation: conditioning on (R_t, S_t, A_t) blocks all paths from S_{t+1} to O_t. For subsequent analysis, we define the extended propensity score of non-missingness $e_t(s, a, r, s') = P(O_t = 1 \mid S_t = s, A_t = a, R_t = r, S_{t+1} = s')$. The choice of S_{t+1} aligns with the recurring idea in POMDPs, i.e., leveraging future states or observations to serve as proxy latent state information (Littman & Sutton, 2001; Singh et al., 2003; Uehara et al., 2023; Xu et al., 2023).

For identification, rather than explicitly modeling the missingness mechanism under MNAR, our goal is to recover the full-data conditional mean reward $\mathbb{E}[R_t \mid S_t, A_t]$ using only observable quantities. Note that under MNAR, the observed reward conditional expectation $\mathbb{E}\big[R_t \mid S_t = s, A_t = a, O_t = 1\big]$ generally differs from the target $\mathbb{E}\big[R_t \mid S_t = s, A_t = a\big]$, and directly using observed rewards would lead to biased policy evaluation.

We adopt a bridge-based imputation strategy for missing rewards motivated by proximal causal inference (Tchetgen Tchetgen et al., 2020; Cui et al., 2024). Related bridge constructions also appear in the literature on confounded POMDPs (Miao et al., 2022; Shi et al., 2022; Hong et al., 2023; Li et al., 2025). The core idea is to construct functions that can learn the missing rewards in an unbiased manner by exploiting the relationship between rewards and next states.

Specifically, we introduce a sequence of bridge functions $\{b_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}\}_{t=1}^{T}$ satisfying the moment condition that

$$\mathbb{E}[b_t(S_t, A_t, S_{t+1}) \mid R_t, S_t, A_t] = R_t, \quad a.s. \tag{2}$$

Equation (2) links the target quantity of interest to the observable offline distribution, and converts the recovery of $\mathbb{E}[R_t \mid S_t, A_t]$ into an estimable conditional moment problem.

Taking conditional expectation of (2) given (S_t, A_t) yields

$$\begin{aligned} &\mathbb{E}[b_t(s, a, S_{t+1}) \mid S_t = s, A_t = a] \\ =&\mathbb{E}(R_t \mid S_t = s, A_t = a) := \bar{r}_t(s, a). \end{aligned} \tag{3}$$

Thus, the bridge function reproduces the correct one-step conditional mean reward required by the Bellman recursion.

Moreover, a crucial observation which enables practical estimation is that

$$P(S_{t+1} \mid R_t, S_t, A_t, O_t = 1) = P(S_{t+1} \mid R_t, S_t, A_t),$$

by Assumption 4.1. This implies that the bridge moment condition in Equation (2) can be identified from the observed subset $\{O_t = 1\}$. This means we can estimate the bridge function b_t using only samples where rewards are observed, and then evaluate it at samples with $O_t = 0$ to impute the missing rewards.

We further introduce the following assumptions for the identification of the policy value.

Assumption 4.3 (Positivity)**.** For all $t = 1, \dots, T$, and for all $(s, a, r) \in \mathcal{S} \times \mathcal{A} \times \mathcal{R}$, $0 < e_t(s, a, r) < 1$.

The positivity assumption ensures that every state-action-reward triple has a positive probability of being both observed and unobserved, which is commonly used in causal inference literature.

Assumption 4.4 (Completeness)**.** For all $(s, a) \in \mathcal{S} \times \mathcal{A}$, $t = 1, \dots, T$,

(1) For any square-integrable function h,

$$\begin{aligned} &\mathbb{E}[h(R_t) \mid S_t = s, A_t = a, S_{t+1}] \\ =&\int h(R_t) p(R_t \mid s, a, S_{t+1}) dR_t = 0, \quad a.s. \end{aligned}$$

if and only if $h(R_t) = 0, \quad a.s.$;

(2) For any square-integrable function g,

$$\begin{aligned} &\mathbb{E}[g(S_{t+1}) \mid R_t, S_t = s, A_t = a] \\ =&\int g(S_{t+1}) p(S_{t+1} \mid R_t, s, a) dS_{t+1} = 0, \quad a.s. \end{aligned}$$

if and only if $g(S_{t+1}) = 0, \quad a.s.$

Assumption 4.4 guarantees the existence and uniqueness of the bridge functions b_t, for $t = 1, \dots, T$. Completeness assumptions are standard in the proximal causal inference (Tchetgen Tchetgen et al., 2020; Cui et al., 2024), where they ensure that the conditional expectations are sufficiently rich to identify the target functional.

To provide concrete intuition, we characterize completeness in the tabular setting.

Example 4.5 (Completeness in tabular setting)**.** *Assume $\mathcal{S}$, $\mathcal{R}$ are tabular. Let matrix $M_{t,s,a} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{R}|}$ where $M_{t,s,a}(s', r) := P(S_{t+1} = s' \mid R_t = r, S_t = s, A_t = a)$, $s' \in \mathcal{S}, r \in \mathcal{R}$. Then*

1. *If for all (t, s, a), $rank(M_{t,s,a}) = |\mathcal{R}|$, then Assumption 4.4 (1) holds, and hence the bridge exists;*
2. *If for all (t, s, a), $rank(M_{t,s,a}) = |\mathcal{S}|$, then Assumption 4.4 (2) holds, and hence the bridge is unique;*
3. *If $|\mathcal{S}| = |\mathcal{R}|$, then $M_{t,s,a}$ is invertible for all (t, s, a) if and only if both Assumption 4.4 (1) and Assumption 4.4 (2) hold.*

Then we give the identification results for policy value as follows:

Theorem 4.6 (Policy value identification)**.** *For an augmented MDP satisfying Assumptions 3.1, 3.2 and 4.1 to 4.4 and some regularity conditions, there always exist bridges $\{b_t\}_{t=1}^T$ that satisfy Equation (2), and the policy value then can be identified using $\{b_t\}_{t=1}^T$.*

See Appendix C for other regularity conditions and proof.

Based on Theorem 4.6, we develop a value-based approach for policy value identification, which circumvents modeling the missing mechanism explicitly, in contrast to approaches such as Miao & Tchetgen Tchetgen (2016); Miao & Tchetgen (2018); Wang et al. (2025). This is practically significant as it can avoid the high variance induced from IPW methods or requires strong parametric assumptions about the missingness. Our identification procedure consists of three steps.

Step 1 (Learn b_t). For each $t = 1, \dots, T$, learn b_t from

$$\mathbb{E}\big[b_t(s, a, S_{t+1}) \mid R_t = r, S_t = s, A_t = a\big] = r,$$

using the observed subset with $O_t = 1$. This step leverages the shadow variable structure to extract reward information from state transitions.

Step 2 (Identify Q^π and V^π). Define the imputed reward

$$\widetilde{R}_t := O_t R_t + (1 - O_t) b_t(S_t, A_t, S_{t+1}), \tag{4}$$

and solve the Bellman recursion

$$\begin{aligned} Q_t^\pi(s, a) =& \mathbb{E}[\widetilde{R}_t + V_{t+1}^\pi(S_{t+1}, O_t) \mid S_t = s, A_t = a], \\ V_t^\pi(s, o_-) =& \sum_a \pi_t(a \mid s, o_-) Q_t^\pi(s, a), \quad V_{T+1}^\pi \equiv 0. \end{aligned}$$

It is trivial to verify that

$$\mathbb{E}[\widetilde{R}_t \mid R_t, S_t, A_t] = R_t, \quad a.s. \tag{5}$$

Step 3 (Identify the policy value). Compute $V(\pi) = \mathbb{E}_{\widetilde{S}_1 \sim \widetilde{\rho}_1}[V_1^\pi(\widetilde{S}_1)]$ by backward induction.

5. Estimation

In this section, we discuss estimation of policy value and propose a FQE-style estimation method. To estimate the policy value, it suffices to estimate the bridge functions $\{b_t\}$ from conditional moment models

$$\mathbb{E}[b_t(S_t, A_t, S_{t+1}) - R_t \mid R_t, S_t, A_t, O_t = 1] = 0, \quad a.s., \tag{6}$$

which can be viewed as nonparametric instrumental variable (NPIV) problems. A natural approach would be to directly minimize the squared conditional moment

$$\min_{b_t \in \mathcal{B}^{(t)}} \mathbb{E}\Big[\big(\mathbb{E}[b_t(S_t, A_t, S_{t+1}) - R_t \mid R_t, S_t, A_t, O_t = 1]\big)^2\Big].$$

However, this is not implementable for a single batch of trajectories because the squared conditional moments can lead to the double-sampling issue (Baird et al., 1995; Sutton et al., 1998). To circumvent the double-sampling problem, we adopt a min-max estimator for b_t (Dikkala et al., 2020). The key insight is to replace the squared conditional moment with a saddle-point formulation that can be estimated from a single batch of samples.

For each time step t, we solve

$$\begin{aligned} \min_{b_t \in \mathcal{B}^{(t)}} \sup_{g \in \mathcal{G}^{(t)}} \frac{1}{n_t} \sum_{i \in \mathcal{I}_t^{\text{obs}}} \big[(b_t(S_{t,i}, A_{t,i}, S_{t+1,i}) - R_{t,i}) g_t(R_{t,i}, \\ S_{t,i}, A_{t,i})\big] - \lambda(\|g_t\|_{\mathcal{G}^{(t)}}^2 + \frac{U}{\delta^2}\|g_t\|_2^2) + \lambda\mu\|b_t\|_{\mathcal{B}^{(t)}}^2, \end{aligned} \tag{7}$$

where $\mathcal{I}_t^{\text{obs}} = \{i \in \{1, \dots, n\} : O_{t,i} = 1\}$ denotes the observed dataset at time t, and $n_t = |\mathcal{I}_t^{\text{obs}}|$. We denote the function classes of g_t and b_t by $\mathcal{G}^{(t)}$, $\mathcal{B}^{(t)}$, which can be chosen as finite dimensional linear spaces, and infinite dimensional spaces like RKHSs, neural networks, etc. We focus on RKHSs in this paper. Let $\mathcal{Q}^{(t)}$ be the RKHS containing function Q_t. The term $\lambda\frac{U}{\delta^2}\|g_t\|_2^2$ is the L_2 penalty on the critic function g_t. The norms $\|\cdot\|_{\mathcal{G}^{(t)}}^2$, $\|\cdot\|_{\mathcal{B}^{(t)}}^2$, $\|\cdot\|_{\mathcal{Q}^{(t)}}^2$ denote the functional norm associated with $\mathcal{G}^{(t)}$, $\mathcal{B}^{(t)}$, $\mathcal{Q}^{(t)}$. $\lambda, U, \delta, \mu > 0$ are tuning parameters for the penalties.

Algorithm 1 Proximal FQE algorithm

Input: Offline dataset $\mathcal{D} = \{\tau_i\}_{i=1}^n$, where $\tau_i = \{(S_{t,i}, A_{t,i}, O_{t,i}, R_{t,i}^{\text{obs}}, S_{t+1,i})\}_{t=1}^T$, target policy $\pi = \{\pi_t\}_{t=1}^T$, horizon T, function classes $\{\mathcal{B}^{(t)}, \mathcal{G}^{(t)}, \mathcal{Q}^{(t)}\}$.
Initialize: $\widehat{V}_{T+1}^{\pi}(\cdot,\cdot) \leftarrow 0$.
for $t = T$ **down to** 1 **do**
  Bridge fitting: Obtain $\hat{b}_t$ by solving Equation (7) on $\mathcal{I}_t^{\text{obs}}$.
  Imputation: for all $i = 1, \ldots, n$ set $\widehat{\widetilde{R}}_{t,i} \leftarrow R_{t,i}^{\text{obs}} + (1 - O_{t,i})\hat{b}_t(S_{t,i}, A_{t,i}, S_{t+1,i})$.
  Targets for Bellman regression:
  if $t < T$ **then**
    $y_{t,i} \leftarrow \widehat{\widetilde{R}}_{t,i} + \widehat{V}_{t+1}^{\pi}(S_{t+1,i}, O_{t,i})$, $i = 1, \ldots, n$.
  else
    $y_{t,i} \leftarrow \widehat{\widetilde{R}}_{t,i}$, $i = 1, \ldots, n$.
  end if
  Fit Q_t**:** regress $y_{t,i}$ on $(S_{t,i}, A_{t,i})$ by Equation (8) to obtain $\widehat{Q}_t$.
  Define V_t^{π}**:** $\widehat{V}_t^{\pi}(s, o_-) \leftarrow \sum_a \pi_t(a \mid s, o_-)\widehat{Q}_t(s, a)$.
end for
Output: $\widehat{V}(\pi) \leftarrow \frac{1}{n}\sum_{i=1}^n \widehat{V}_1^{\pi}(S_{1,i}, 0)$.

Then, we can substitute the estimates into fitted-Q-evaluation (FQE) algorithm and obtain the estimate of policy value $\widehat{V}(\pi)$. See Algorithm 1 for the point-estimated policy value estimation algorithm, where we use penalized nonparametric least squares to learn Q_t:

$$\widehat{Q}_t = \arg\min_{f \in \mathcal{Q}^{(t)}} \frac{1}{n}\sum_{i=1}^n \big(f(S_{t,i}, A_{t,i}) - y_{t,i}\big)^2 + \lambda_t \|f\|_{\mathcal{Q}^{(t)}}^2, \tag{8}$$

where $y_{t,i}$ is defined in Algorithm 1.

6. Theoretical results

In this section, we establish consistency and finite-sample estimation error bounds for bridges $\hat{b}_t$ and the policy value.

6.1. Preliminaries

Definition 6.1 (Local Rademacher Complexity (Bartlett et al., 2005))**.** For any function class $\mathcal{G}$ defined over random variable X and radius $\delta > 0$, the *local Rademacher complexity* is

$$\mathcal{R}_n(\mathcal{G}, \delta) = \mathbb{E}_{\varepsilon, X}\Big[\sup_{g \in \mathcal{G}: \|g\|_2 \le \delta} \Big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i)\Big|\Big],$$

where $\{X_i\}$ are i.i.d. samples of X and $\{\varepsilon_i\}$ are Rademacher random variables. $\|g\|_2^2 := \mathbb{E}[g(X)^2]$ is the L_2 norm of function g.

Suppose the function class $\mathcal{G}$ satisfies

1. *symmetric*, if $g \in \mathcal{G}$ then $-g \in \mathcal{G}$;
2. *star-shaped*, if $g \in \mathcal{G}$ then $rg \in \mathcal{G}$ for all $r \in [0, 1]$;
3. *b-uniformly bounded*, $\|g\|_\infty := \sup_{x \in \mathcal{X}} |g(x)| \le b$ for all $g \in \mathcal{G}$.

Then, the critical radius of such function class $\mathcal{G}$, denoted by δ_n, is the smallest solution to the inequality $\mathcal{R}_n(\mathcal{G}, \delta) \le \frac{\delta^2}{b}$.

6.2. Bridge function estimation error bound

For notational simplicity, define the projection operator $\mathcal{T}_t : \mathcal{L}^2\{\mathcal{S} \times \mathcal{A} \times \mathcal{S}\} \to \mathcal{L}^2\{\mathcal{R} \times \mathcal{S} \times \mathcal{A}\}$, which satisfies

$$\mathcal{T}_t b_t = \mathbb{E}[b_t(S_t, A_t, S_{t+1}) | R_t, S_t, A_t].$$

Assumption 6.2 (Boundedness of $\mathcal{T}_t$)**.** For any $b_t \in \mathcal{B}^{(t)}$, $\mathcal{T}_t b_t \in \mathcal{G}^{(t)}$, and there exists $L > 0$ such that

$$\|\mathcal{T}_t b_t\|_{\mathcal{G}^{(t)}} \le L \|b_t\|_{\mathcal{B}^{(t)}}.$$

Assumption 6.3 (Realizability)**.** Suppose the true bridge function b_t^* lies in function class $\mathcal{B}^{(t)}$. Similarly, we also assume $Q_t^{\pi} \in \mathcal{Q}^{(t)}$.

In practice, the boundedness assumption often holds when the conditional distribution of S_{t+1} given (R_t, S_t, A_t) is sufficiently smooth. Also, Realizability is a standard assumption in nonparametric estimation, requiring that the function classes are rich enough to contain the true targets.

Next, we define the B_t-bounded norm subset of $\mathcal{B}^{(t)}$ as $\mathcal{B}_B^{(t)} := \{b_t \in \mathcal{B}^{(t)} : \|b_t\|_{\mathcal{B}^{(t)}} \le B_t\}$ and U_t-bounded norm subset of $\mathcal{G}^{(t)}$ as $\mathcal{G}_U^{(t)} := \{g_t \in \mathcal{G}^{(t)} : \|g_t\|_{\mathcal{G}^{(t)}} \le U_t\}$.

Assumption 6.4 (Richness of test function class)**.** We suppose the test function approximation error within subset $\mathcal{G}^{(t)}_{L^2\|(b-b_t^*)\|^2_{\mathcal{B}^{(t)}}}$ is bounded by

$$\sup_{b \in \mathcal{B}_B^{(t)}} \inf_{g_t \in \mathcal{G}^{(t)}_{L^2\|b-b_t^*\|^2_{\mathcal{B}^{(t)}}}} \|g_t - \mathcal{T}_t(b - b_t^*)\|_2 \le \eta_t < \infty.$$

This shows that function class $\mathcal{G}^{(t)}$ is rich enough so that $\mathcal{T}_t(b - b_t^*)$ admits an L_2-approximation within $\mathcal{G}^{(t)}_{L^2\|(b-b_t^*)\|^2_{\mathcal{B}^{(t)}}}$ uniformly over $b \in \mathcal{B}_B^{(t)}$.

Now we are ready to analyze min-max estimator $\hat{b}_t$ estimated by Equation (7).

Theorem 6.5 (Projected bridge estimation error bound (Dikkala et al., 2020; Miao et al., 2022))**.** *For any $t = 1, \ldots, T$, suppose function class $\mathcal{G}^{(t)}$ is star-shaped and symmetric. Suppose $\mathcal{G}^{(t)}$ and $\mathcal{B}^{(t)}$ are 1-uniformly bounded. Define product class*

$$\mathcal{J}_{B,U}^{(t)} := \Big\{((s,a,s'),(r,s,a)) \mapsto \alpha(b_t(s,a,s') - b_t^*(s,a,s'))g_{b,t}^U(r,s,a) \mid b_t - b_t^* \in \mathcal{B}_B^{(t)}, \alpha \in [0,1]\Big\},$$

where $g^U_{b,t} = \arg\min_{g_t \in \mathcal{G}^{(t)}_U} \|g_t - \mathcal{T}_t(b - b^*_t)\|_2$. *Define the two critical radii for function class* $\mathcal{G}^{(t)}_{3U}$ *and* $\mathcal{J}^{(t)}_{B,L^2B}$, *namely* $\delta^{\mathcal{G}}_t$ *and* $\delta^{\mathcal{J}}_t$, *and their maximum* $\delta_{n_t} := \max\{\delta^{\mathcal{G}}_t, \delta^{\mathcal{J}}_t\}$. *Let* $\delta_t = \delta_{n_t} + c_0\sqrt{\frac{\log(c_1/\zeta)}{n_t}}$. *Assume* $\eta_t \lesssim \delta_t$, *then if* $\lambda \asymp \frac{\delta_t^2}{U}$ *and* $\mu \geq \frac{4}{3}L^2 + \frac{36}{B_t\lambda}\delta_t^2$, *we have with probability* $1 - 3\zeta$, *the following bound holds*

$$\left\|\mathcal{T}_t(\hat{b}_t - b^*_t)\right\|_2 \lesssim \delta_t \max\{1, \|b^*_t\|_{\mathcal{B}^{(t)}}\}.$$

Corollary 6.6 (RKHS cases with polynomial eigen decay)**.** *We further suppose* $\mathcal{B}^{(t)}$ *and* $\mathcal{G}^{(t)}$ *are RKHSs for all* $t = 1, \ldots, T$. $K_{B,t}$ *and* $K_{G,t}$ *are the kernels of* $\mathcal{B}^{(t)}$ *and* $\mathcal{G}^{(t)}$ *with non-increasing eigenvalues* $\{\lambda^{\mathcal{B}}_{t,j}\}_{j=1}^{\infty}$ *and* $\{\lambda^{\mathcal{G}}_{t,j}\}_{j=1}^{\infty}$. *We assume polynomial decay on eigenvalues, i.e., for some* $\alpha_B, \alpha_G > \frac{1}{2}$, $\lambda^{\mathcal{B}}_{t,j} \lesssim j^{-2\alpha_B}$, $\lambda^{\mathcal{G}}_{t,j} \lesssim j^{-2\alpha_G}$, $j \to \infty$. *Let* $\alpha_{\min} := \min\{\alpha_B, \alpha_G\}$. *Then the critical radius in Theorem 6.5 satisfy* $\delta_{n_t} \lesssim \max\{\sqrt{U_t}, LB_t\} n_t^{-\frac{\alpha_{\min}}{2\alpha_{\min}+1}} \log n_t$. *Consequently, under the conditions of Theorem 6.5, for all* $t = 1, \ldots, T$, *with probability at least* $1 - 3\zeta$,

$$\|\mathcal{T}_t(\hat{b}_t - b^*_t)\|_2 \lesssim \sqrt{\log(c_1/\zeta)} n_t^{-\frac{\alpha_{\min}}{2\alpha_{\min}+1}} \log n_t.$$

Theorem 6.5 provides a finite-sample bound for the projected error $\|\mathcal{T}_t(\hat{b}_t - b^*_t)\|_2$, where linear operator $\mathcal{T}_t$ maps a bridge function to a conditional expectation given (R_t, S_t, A_t). However, for downstream analysis we also need control of the rooted mean-squared error (RMSE) $\|\hat{b}_t - b^*_t\|_2$.

In general, converting projected error bounds into L_2 error bounds is nontrivial because conditional moment problems are typically ill-posed inverse problems: the operator $\mathcal{T}_t$ is often compact and hence may not admit a stable inverse on an unrestricted function class. This phenomenon and the role of regularization in such conditional moment models are well-studied in the semi-/nonparametric literature; see, e.g., (Chen & Reiss, 2011; Chen & Pouzo, 2012). We introduce an ill-posedness measure for the conditional expectation operator $\mathcal{T}_t$, following the definition in Dikkala et al. (2020). Since the true bridge b^*_t is not assumed to lie in $\mathcal{B}^{(t)}_B$, we define the best approximation within the ball

$$b_{t,*} := \arg\min_{b \in \mathcal{B}^{(t)}_B} \|b - b^*_t\|_2, \quad \varepsilon_t(B_t) := \inf_{b \in \mathcal{B}^{(t)}_B} \|b - b^*_t\|_2.$$

Definition 6.7 (Measure of ill-posedness)**.** Define the ill-posedness coefficient $\tau_t(B_t) := \sup_{b \in \mathcal{B}^{(t)}_B} \frac{\|b - b_{t,*}\|_2}{\|\mathcal{T}_t(b - b_{t,*})\|_2}$, and assume $\tau_t(B_t) < \infty$.

By combining Theorem 6.5 with Definition 6.7, we obtain an L_2 error bound for the bridge estimator:

$$\|\hat{b}_t - b^*_t\|_2 \leq \tau_t(B_t)\delta_t + (\tau_t(B_t) + 1)\varepsilon_t(B_t).$$

The choice of B_t trades off the approximation bias $\varepsilon_t(B_t)$ and the ill-posedness factor $\tau_t(B_t)$. Consider the whole function class $\mathcal{B}^{(t)}$, where $b_{t,*} = b^*_t$ and $\varepsilon_t(B_t) = 0$ under Assumption 6.3, this gives the global ill-posedness

$$\tau_t := \sup_{b \in \mathcal{B}^{(t)}} \frac{\|b - b_{t,*}\|_2}{\|\mathcal{T}_t(b - b_{t,*})\|_2},$$

where we assume $\tau_t < \infty$. The RMSE is given by $\|\hat{b}_t - b^*_t\|_2 \lesssim \tau_t\delta_t$.

Theorem 6.8 (Bridge estimation error bound)**.** *For any* $t = 1, \ldots, T$, *suppose function class* $\mathcal{G}^{(t)}$ *is star-shaped and symmetric. Suppose* $\mathcal{G}^{(t)}$ *and* $\mathcal{B}^{(t)}$ *are* 1*-uniformly bounded. Consider min-max estimator* $\hat{b}_t$ *estimated by Equation (7). Define function classes* $\operatorname{star}\left(\mathcal{B}^{(t)} - b^*_t\right) = \{r(b - b^*_t) : b - b^*_t \in \mathcal{B}^{(t)}_B, r \in [0,1]\}$, *and* $\operatorname{star}\left(\mathcal{T}_t(\mathcal{B}^{(t)} - b^*_t)\right) = \{r\, g^U_{b,t} : b - b^*_t \in \mathcal{B}^{(t)}_B, r \in [0,1]\}$, *where* $g^U_{b,t} = \arg\min_{g_t \in \mathcal{G}^{(t)}_U} \|g_t - \mathcal{T}_t(b - b^*_t)\|_2$. *Define the* δ_{n_t} *as the upper bound on the critical radii of* $\mathcal{G}^{(t)}_{3U}$ *and the two function classes. Let* $\delta_t = \delta_{n_t} + c_0\sqrt{\frac{\log(c_1/\zeta)}{n_t}}$. *Assume* $\eta_t \lesssim \delta_t$, *then if* $\lambda \asymp \frac{\delta_t^2}{U}$ *and* $\mu \geq \frac{4}{3}L^2 + \frac{36}{B_t\lambda}\delta_t^2$, *then with probability* $1 - 3\zeta$, *the following bound holds*

$$\left\|\hat{b}_t - b^*_t\right\|_2 \lesssim \tau_t\delta_t \max\{1, \|b^*_t\|_{\mathcal{B}^{(t)}}\}.$$

6.3. Policy value estimation error bound

Based on Theorem 6.8, we can further bound the OPE error of the policy value $\widehat{V}(\pi)$ estimated from $Algorithm\ 1$.

Theorem 6.9 (Policy value estimation error bound)**.** *Suppose RKHSs* $\mathcal{Q}^{(t)}$, $\mathcal{B}^{(t)}$, $\mathcal{G}^{(t)}$ *have polynomial eigen-decay rate*

$$\lambda^{\mathcal{Q}}_{t,j} \lesssim j^{-2\alpha_Q}, \lambda^{\mathcal{B}}_{t,j} \lesssim j^{-2\alpha_B}, \lambda^{\mathcal{G}}_{t,j} \lesssim j^{-2\alpha_G},$$

where $\alpha_Q, \alpha_B, \alpha_G > 1/2$. *Define* $\alpha_{\min} = \min\{\alpha_Q, \alpha_B, \alpha_G\} > \frac{1}{2}$. *Denote* $\delta_{t,*} = \bar{\delta}_{t,*} + c_0\sqrt{\frac{\log(c_1 T/\zeta)}{n}}$ *for some* $c_0, c_1 > 0$ *where* $\bar{\delta}_{t,*}$ *is the upper bound of the critical radii of difference classes* $\Delta\mathcal{Q}^{(t)}$, $\Delta\mathcal{Q}^{(t+1)}$, $\Delta\mathcal{B}^{(t)}$ *and* $\mathcal{G}^{(t)}_U$ *defined in Appendix F. Suppose* $\lambda_t \asymp (\delta_{\Delta^{(t)}_{\mathcal{Q}}})^2$ *and let* $\tau_{\max} = \max_{t \leq T} \tau_t$. *Under Assumptions 3.3, 3.4, 6.2 and 6.3 and assumptions for Theorem 4.6, with probability at least* $1 - \zeta$, *the policy value estimation error is bounded by*

$$\left|\widehat{V}(\pi) - V(\pi)\right| \lesssim K\tau_{\max}T^2\sqrt{\log(c_1 T/\zeta)} n^{-\frac{\alpha_{\min}}{2\alpha_{\min}+1}} \log n.$$

Without considering the ill-posedness, our OPE error bound achieves the optimal rate in n in the classical nonparametric regression (Stone, 1982). Our error bound exhibits a T^2 dependence on the horizon, which arises from error propagation through the Bellman recursion and the additional complexity introduced by estimating bridge functions under the

MNAR setting. For comparison, Wang et al. (2024) provide a fine-grained analysis of FQE under fully observed rewards. Under the completeness assumption for Q-functions alone, they establish an error bound of order $\mathcal{O}(T^{1.5}\sqrt{1/n})$ for both parametric and nonparametric settings, improving upon the $\mathcal{O}(T^2\sqrt{\kappa/n})$ bounds in prior work (Duan et al., 2020; Zhang et al., 2022). With an additional realizability assumption on the probability ratio functions w_t^π, the rate further improves to $\mathcal{O}(T\tilde{\kappa}\sqrt{1/n})$, matching the sharpest known bound under the tabular setting (Yin & Wang, 2020). The additional T factor in our bound compared to the $T^{1.5}$ rate in Wang et al. (2024) partially reflects the cost of correcting for MNAR rewards through the bridge function mechanism.

7. Experiments

In this section, we conduct simulation studies to evaluate the performance of the proposed OPE estimator with rewards MNAR. We compare it with an IPW based OPE method (Wang et al., 2025) and a naive FQE baseline in finite-horizon episodic MDPs. We include the naive FQE estimator as a simple baseline that ignores the MNAR mechanism and an IPW based estimator obtained by adapting the weighting scheme of (Wang et al., 2025), which was originally developed for a trajectory dropout model. Our code is available at `https://anonymous.4open.science/r/OPE-MDP-MNAR-4A88/`.

We set state $S_t = (S_{t,1}, S_{t,2})^\top \in \mathcal{S} = \mathbb{R}^2$ as a two-dimensional vector. The action space is binary, $\mathcal{A} = \{-1, 1\}$. Let $\mathcal{O} = \{0, 1\}$, $\mathcal{R} = \mathbb{R}$. The propensity score is set as $e_t(S_t, A_t, R_t) = \text{expit}\left(1 - 0.1A_t + 0.2(1, -2)^\top S_t + 2.5R_t\right)$. The target policy we want to evaluate is given by

$$P_\pi(A_t = 1 \mid S_t, O_{t-1}) = \text{expit}\left\{3[(1, 0.3)^\top S_t + 0.5 - 0.8(2O_{t-1} - 1)]\right\}$$

See Figure 3 in Appendix B for visualization of the generated data.

For function classes, we choose Gaussian kernels for $\mathcal{G}^{(t)}$ and $\mathcal{B}^{(t)}$, and consider mean absolute error (MAE) and mean squared error (MSE) as evaluation metrics.

We conduct two sets of experiments. In the first, we fix the horizon at $T = 8$ and vary the sample size $n \in \{128, 256, 512, 1024, 2048\}$. In the second, we fix the sample size at $n = 512$ and vary the horizon $T \in \{2, 4, 8, 16, 32\}$. For each configuration, we repeat the experiment over five random seeds and report the MAE and MSE of the estimated policy value. Figure 2 shows the MAE of `prox` (our method), and two baselines `naive` (naive FQE) and `ipw` (IPW-FQE). The left (MAE vs n) shows that `prox` achieves the fastest convergence rate as the sample size increases, and consistently attains the smallest error across all n, while `ipw` remains an order of magnitude higher. `naive` decreases only slowly with n since ignores the MNAR mechanism and regresses on observed rewards, which typically incurs a bias that does not vanish quickly with more data. In the right figure (MAE vs T), `ipw` baseline becomes markedly unstable for larger T, with a sharp increase at $t = 32$, indicating variance blow-up induced by inverse propensity weights. In contrast, `prox` remains the most stable across horizons, showing a significantly slower growth in MAE as T increases. `naive` baseline is consistently worse, reflecting the bias introduced by ignoring reward MNAR. Overall, the convergence rate is consistent with our theoretical results. Additional simulation details can be found in Appendix B.

8. Conclusion and Discussion

We study OPE in MDPs with MNAR rewards, proposing a bridge function approach that recovers the conditional mean reward without explicitly modeling the missingness mechanism. Unlike IPW-based methods (Wang et al., 2025) that require external auxiliary variables, our method uses the next state as an endogenous shadow variable, avoiding additional data requirements and IPW variance inflation.

A limitation of our framework is that the reward missingness process may still be influenced by unobserved confounding factors that are not fully captured by the observed trajectories. While our identification relies on assumptions through the bridge function, violations of these assumptions may lead to biased policy value estimates in practice.

To address this concern, an important future direction is to incorporate sensitivity analysis into the proposed OPE framework. By parameterizing deviations from the bridge moment conditions or the completeness assumptions, one can assess the robustness of policy value estimates to potential unobserved confounding in the missingness mechanism. Such sensitivity analyses have been studied in proximal causal inference and POMDPs, and adapting their frameworks from proximal causal inference and POMDPs to the reward-missingness mechanism in MDPs is an important direction for future work.

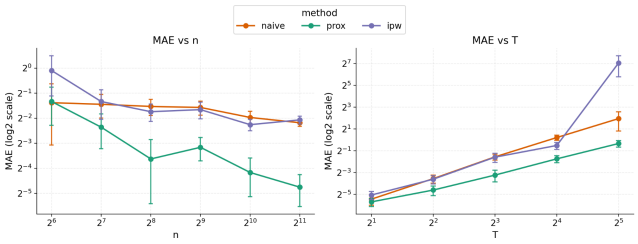

Figure 2. Policy value estimation MAE versus sample size and horizon. Left: MAE as a function of n with $T = 8$. Right: MAE as a function of T with $n = 512$.

Impact Statement

This paper presents work whose goal is to advance the field of machine learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Baird, L. et al. Residual algorithms: Reinforcement learning with function approximation. In *Proceedings of the twelfth international conference on machine learning*, pp. 30–37, 1995.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. 2005.

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

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

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International conference on machine learning*, pp. 1042–1051. PMLR, 2019.

Chen, X. and Pouzo, D. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 2012.

Chen, X. and Reiss, M. On rate optimality for ill-posed inverse problems in econometrics. *Econometric Theory*, 27(3):497–521, 2011.

Chu, J., Yang, S., and Lu, W. Multiply robust off-policy evaluation and learning under truncation by death. In *International Conference on Machine Learning*, pp. 6195–6227. PMLR, 2023.

Cui, Y., Pu, H., Shi, X., Miao, W., and Tchetgen Tchetgen, E. Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359, 2024.

Devlin, S. and Kudenko, D. Dynamic potential-based reward shaping. *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, pp. 433–440, 2012.

Dikkala, N., Lewis, G., Mackey, L., and Syrgkanis, V. Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems*, 33:12248–12262, 2020.

Duan, Y., Jia, Z., and Wang, M. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pp. 2701–2709. PMLR, 2020.

Duan, Y., Jin, C., and Li, Z. Risk bounds and rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 2892–2902. PMLR, 2021.

Enders, C. K. *Applied missing data analysis*. Guilford Publications, 2022.

Fischer, S. and Steinwart, I. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.

Foster, D. J. and Syrgkanis, V. Orthogonal statistical learning. *The Annals of Statistics*, 51(3):879–908, 2023.

Fukumizu, K., Gretton, A., Lanckriet, G., Schölkopf, B., and Sriperumbudur, B. K. Kernel choice and classifiability for rkhs embeddings of probability distributions. *Advances in neural information processing systems*, 22, 2009.

Hong, M., Qi, Z., and Xu, Y. A policy gradient method for confounded pomdps. *arXiv preprint arXiv:2305.17083*, 2023.

Jaakkola, T., Singh, S. P., and Jordan, M. I. Reinforcement learning algorithm for partially observable markov decision problems. *Advances in Neural Information Processing Systems*, pp. 345–352, 1995.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. *Planning and acting in partially observable stochastic domains*, volume 101. Elsevier, 1998.

Kallus, N. and Uehara, M. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. In *International Conference on Machine Learning*, pp. 5078–5088. PMLR, 2020.

Kallus, N. and Zhou, A. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Advances in Neural Information Processing Systems*, 33:22293–22304, 2020.

Kress, R. *Linear integral equations*, volume 82. Springer, 1989.

Krieg, D. Tensor power sequences and the approximation of tensor product operators. *Journal of Complexity*, 44: 30–51, 2018.

Le, H. M., Voloshin, C., and Yue, Y. Batch policy learning under constraints. *International Conference on Machine Learning*, pp. 3703–3712, 2019.

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

Li, Y., Han, E., Hu, Y., Zhou, W., Qi, Z., Cui, Y., and Zhu, R. Reinforcement learning with continuous actions under unmeasured confounding. *Journal of the American Statistical Association*, (just-accepted):1–26, 2025.

Little, R. J. and Rubin, D. B. Statistical analysis with missing data. *John Wiley & Sons*, 2019.

Littman, M. and Sutton, R. S. Predictive representations of state. *Advances in neural information processing systems*, 14, 2001.

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

Mandel, T., Liu, Y.-E., Levine, S., Brunskill, E., and Popovic, Z. Offline policy evaluation across representations with applications to educational games. *Proceedings of AAMAS*, 2014.

Miao, R., Qi, Z., and Zhang, X. Off-policy evaluation for episodic partially observable markov decision processes under non-parametric models. *Advances in Neural Information Processing Systems*, 35:593–606, 2022.

Miao, W. and Tchetgen, E. T. Identification and inference with nonignorable missing covariate data. *Statistica Sinica*, 28(4):2049, 2018.

Miao, W. and Tchetgen Tchetgen, E. J. On varieties of doubly robust estimators under missingness not at random with a shadow variable. *Biometrika*, 103(2):475–482, 2016.

Miao, W., Tchetgen Tchetgen, E., and Geng, Z. Identification and doubly robust estimation of data missing not at random with an ancillary variable. 2015.

Mohan, K. and Pearl, J. Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037, 2021.

Munos, R. Error bounds for approximate policy iteration. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, pp. 560–567, 2003.

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

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

Ng, A. Y., Harada, D., and Russell, S. Policy invariance under reward transformations: Theory and application to reward shaping. In *International Conference on Machine Learning*, pp. 278–287, 1999.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Park, S., Lu, W., and Yang, S. Evaluating and learning optimal dynamic treatment regimes under truncation by death. *arXiv preprint arXiv:2510.07501*, 2025.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36: 53728–53741, 2023.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. In *Advances in Neural Information Processing Systems*, volume 34, pp. 11702–11716, 2021.

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

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, pp. 31335–31385. PMLR, 2023.

Singh, S. P., Littman, M. L., Jong, N. K., Pardoe, D., and Stone, P. Learning predictive state representations. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 712–719, 2003.

Stone, C. J. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pp. 1040–1053, 1982.

Sun, B. and Tchetgen Tchetgen, E. J. Semiparametric estimation with data missing not at random using an instrumental variable. *Statistica Sinica*, 28(4):1965, 2018.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning: An Introduction*. MIT Press, 2nd edition, 2018.

Sutton, R. S., Barto, A. G., et al. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

Tchetgen Tchetgen, E. J., Ying, A., Cui, Y., Shi, X., and Miao, W. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Tennenholtz, G., Shalit, U., and Mannor, S. Off-policy evaluation in partially observable environments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(06): 10276–10283, 2020.

Tsiatis, A. A., Davidian, M., Holloway, S. T., and Laber, E. B. *Dynamic Treatment Regimes: Statistical Methods for Precision Medicine*. CRC press, 2019.

Uehara, M., Kiyohara, H., Bennett, A., Chernozhukov, V., Jiang, N., Kallus, N., Shi, C., and Sun, W. Future-dependent value-based off-policy evaluation in pomdps. *Advances in Neural Information Processing Systems*, 35, 2022a.

Uehara, M., Shi, C., and Kallus, N. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022b.

Uehara, M., Kiyohara, H., Bennett, A., Chernozhukov, V., Jiang, N., Kallus, N., Shi, C., and Sun, W. Future-dependent value-based off-policy evaluation in pomdps. *Advances in neural information processing systems*, 36: 15991–16008, 2023.

Voloshin, C., Le, H. M., Jiang, N., and Yue, Y. Empirical study of off-policy policy evaluation for reinforcement learning. *Advances in Neural Information Processing Systems*, 2021.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Wang, H., Xu, Y., Lu, W., and Song, R. Off-policy evaluation under nonignorable missing data. *arXiv preprint arXiv:2507.06961*, 2025.

Wang, J., Qi, Z., and Wong, R. K. A fine-grained analysis of fitted q-evaluation: beyond parametric models. *arXiv preprint arXiv:2406.10438*, 2024.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6683–6694, 2021.

Xu, Y., Zhu, J., Shi, C., Luo, S., and Song, R. An instrumental variable approach to confounded off-policy evaluation. In *International Conference on Machine Learning*, pp. 38848–38880. PMLR, 2023.

Yang, L., Cui, Y., Xuan, Y., Wang, C., Belongie, S., and Estrin, D. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM conference on recommender systems*, pp. 279–287, 2018.

Yin, M. and Wang, Y.-X. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 3948–3958. PMLR, 2020.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. D. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Zhang, R., Dai, B., Li, L., and Schuurmans, D. Off-policy fitted q-evaluation with differentiable function approximators: Z-estimation and inference theory. *arXiv preprint arXiv:2202.04970*, 2022.

Zhao, P., Yang, S., and Kim, J. K. Semiparametric inverse propensity weighting for nonignorable missing data. *Biometrika*, 102(3):589–601, 2015.

A. Discussion on shadow variables

In the causal inference literature on missing data, identification typically requires additional assumptions, most commonly instantiated through either instrumental variables or shadow variables. Shadow-variable approaches (Miao et al., 2015; Miao & Tchetgen Tchetgen, 2016) leverage a fully observed variable that is informative about the outcome while being conditionally independent of the missingness mechanism given covariates and the (possibly unobserved) outcome. In contrast, instrumental-variable approaches (Sun & Tchetgen Tchetgen, 2018) posit a variable that shifts the missingness mechanism but has no direct effect on the outcome.

In sequential decision making, Wang et al. (2025) develop an approach that requires specifying a stage-wise shadow variable Z_t that satisfies conditional independence with dropout given $(S_t, A_t, R_{t+1}, S_{t+1})$ while remaining informative about (R_{t+1}, S_{t+1}) on the observed subset, effectively providing the identifying leverage needed to learn the dropout model.

While such a choice can be plausible, it may rely on additional measurements or domain knowledge to select a valid Z_t. In contrast, we adopt an endogenous choice of shadow variable, which is the next state S_{t+1}. Under the exclusion restriction and relevance condition in Assumptions 4.1 and 4.2, S_{t+1} provides a readily available proxy for the MNAR reward without introducing an extra auxiliary variable.

Moreover, a natural extension of S_{t+1} is a multi-step future variable, or a low-dimensional summary thereof. This aligns with a broader predictive-state perspective in POMDPs, where future observations are used to encode information about the latent state (Littman & Sutton, 2001; Singh et al., 2003). More recently, Xu et al. (2023); Uehara et al. (2023) use multi-step futures to stand in for the unobserved state: instead of conditioning on the latent state, they condition on a future window and learn quantities from it, using the future window as a proxy that carries latent-state information.

In missing data problems in MDPs, using longer futures can carry richer information about the missing rewards, and may make relevance and completeness-type conditions more plausible, but it also increases the statistical and computational burden as the future window grows. In practice, these tradeoffs motivate using compact summaries of multi-step futures.

B. Additional Experiment Details

B.1. Additional simulation setups

We set the behavior policy that generates the offline trajectories as

$$P_{\pi^b}(A_t = 1 \mid S_t) = \operatorname{expit}\big(0.3 + (0.8, -0.3)^\top S_t\big).$$

The next state S_{t+1} is generated by transition kernel $S_{t+1} = 0.9S_t + 0.2A_t\mathbf{1}_2 + \mathcal{N}(0, 0.1^2 I_2)$ where $\mathbf{1}_2 = (1,1)^\top$ and initial state $S_1 \sim \mathcal{N}(0, I_2)$. The reward model is

$$R_t = \operatorname{expit}\big[(0.9 - 0.6A_t, -0.7)^\top S_t + (1.3, 2)^\top S_{t+1} - 0.4A_t\big] + \mathrm{Unif}[-0.1, 0.1].$$

The true policy value is estimated through Monte Carlo by averaging over 5000 independent trajectories generated under the target policy. We visualize the generated data when $n = 1000$, $T = 10$ and random seed is 44. See Figure 3 for an overview of the data.

For all the RKHSs, the bandwidths are selected by median heuristic trick (Fukumizu et al., 2009); parameter δ is set to $\delta_t = 5n_t^{-0.4}$ according to (Dikkala et al., 2020). The penalty parameter λ_{rkhs} is chosen by cross-validation.

B.2. Baseline methods

B.2.1. NAIVE FQE

In naive FQE baseline, we ignore the missingness mechanism, and perform FQE only on observed samples:

$$\widehat{Q}_t^{\mathrm{naive}} = \arg\min_{Q \in \mathcal{Q}^{(t)}} \frac{1}{n_t} \sum_{i \in \mathcal{I}_t^{\mathrm{obs}}} \Big(Q(S_{t,i}, A_{t,i}) - y_{t,i}^{\mathrm{naive}}\Big)^2 + \lambda_{\mathrm{rkhs}} \|Q\|^2_{\mathcal{Q}^{(t)}},$$

where

$$y_{t,i}^{\mathrm{naive}} = \begin{cases} R_{t,i}, & t = T, \\ R_{t,i} + \sum_{a \in \mathcal{A}} \pi(a \mid S_{t+1,i}, O_{t,i})\, \widehat{Q}_{t+1}^{\mathrm{naive}}(S_{t+1,i}, a), & t < T. \end{cases}$$

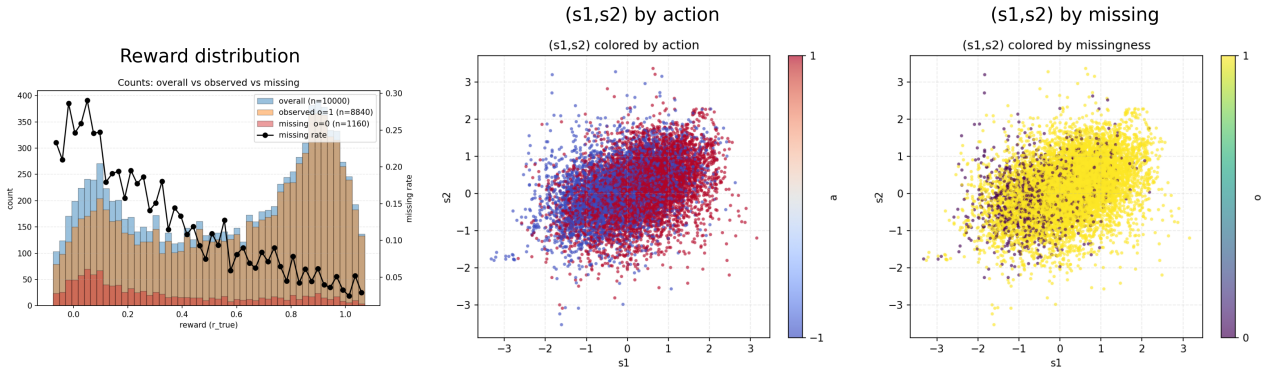

Figure 3. **Data overview. Left**: histogram of the true reward r_{true} with three overlays: overall (blue), observed $O{=}1$ (orange), and missing $O{=}0$ (red). The total missing rate is 11.60%. Missing mass is relatively larger in the low–reward region. **Middle**: state value $S_t = (s_1, s_2)$ colored by action (blue: $a{=}-1$, red: $a{=}+1$) according to target policy, showing both actions across the state space without obvious coverage gaps. **Right**: state value $S_t = (s_1, s_2)$ colored by missingness (yellow: $O{=}1$, purple: $O{=}0$); the non-uniform placement of missing points indicates observation probability varies with state.

As discussed in Section 4, under reward MNAR we generally have

$$\mathbb{E}[R_t \mid S_t = s, A_t = a, O_t = 1] \neq \mathbb{E}[R_t \mid S_t = s, A_t = a] \quad \text{for some } (s, a),$$

so naive FQE, which regresses on observed rewards, can be biased.

B.2.2. IPW-FQE

In IPW-FQE baseline, since Wang et al. (2025) study a dropout model, we modify their method to fit our reward missingness setting. At each stage t we solve

$$\widehat{Q}_t^{\text{ipw}} = \arg\min_{Q \in \mathcal{Q}^{(t)}} \frac{1}{n_t} \sum_{i \in \mathcal{I}_t^{\text{obs}}} w_{t,i} \Big(Q(S_{t,i}, A_{t,i}) - y_{t,i}^{\text{ipw}} \Big)^2 + \lambda_{\text{rkhs}} \|Q\|_{\mathcal{Q}^{(t)}}^2,$$

where

$$y_{t,i}^{\text{ipw}} = \begin{cases} R_{t,i}, & t = T, \\ R_{t,i} + \sum_{a \in \mathcal{A}} \pi(a \mid S_{t+1,i}, O_{t,i})\, \widehat{Q}_{t+1}^{\text{ipw}}(S_{t+1,i}, a), & t < T, \end{cases}$$

and weights $w_{t,i} = \frac{1}{\hat{e}_{t,i}}, \quad i \in \mathcal{I}_t$. The extended propensity score e_t is estimated by logistic regression

$$\hat{e}_t(s, a, \widehat{b}) = P(O_t = 1 \mid S_t = s, A_t = a, \hat{b}_t = \hat{b}).$$

Figure 4 summarizes how the MSE varies with the sample size n (left) and the horizon T (right).

C. Proof of Theorem 4.6

In this section we provide the proof of identification result in Theorem 4.6.

Part A: Identification by bridge functions

We first show that the policy value can be identified by the bridge functions $\{b_t\}_{t=1}^T$ if the bridges exist.

Fix $t \in \{1, \ldots, T\}$ and suppose there exists a measurable function $b_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ satisfying the Equation (2). By Assumptions 4.1 and 4.3, conditioning on the event $O_t = 1$ is well-defined and

$$\mathbb{E}[b_t(S_t, A_t, S_{t+1}) \mid R_t, S_t, A_t, O_t = 1] = \mathbb{E}[b_t(S_t, A_t, S_{t+1}) \mid R_t, S_t, A_t].$$

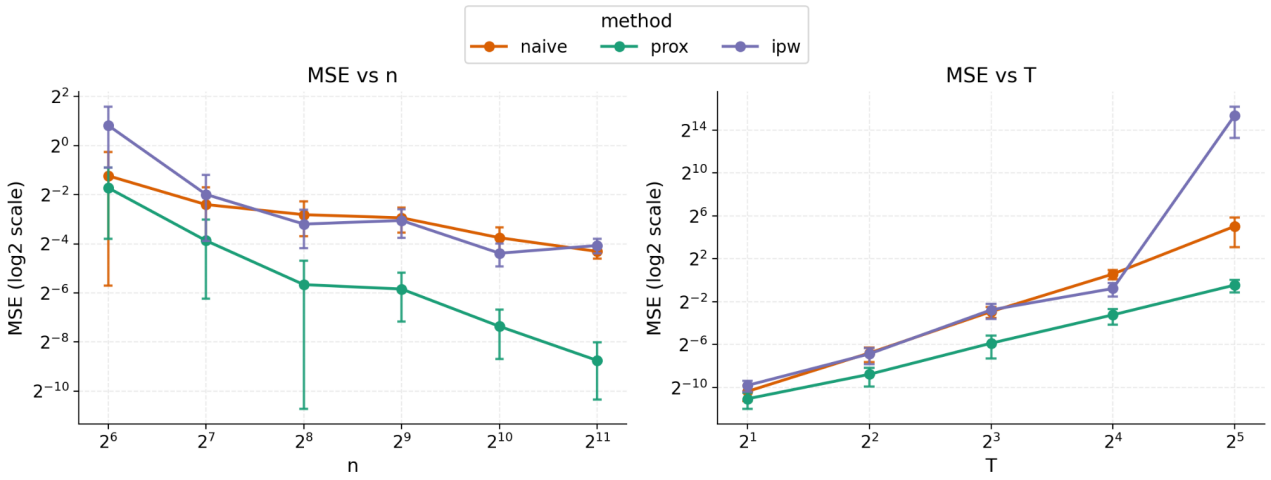

Figure 4. **Policy value estimation MSE versus sample size and horizon. Left (MSE vs n, log scale):** *prox* attains the smallest and most stable MSE; *naive* and *ipw* exhibit substantially slower error decay as n increases. **Right (MSE vs T, log scale):** MSE climbs roughly monotonically with T; *prox* preserves a gap over *naive* at every T, with the gap growing as T increases. *ipw* becomes unstable when T is high.

So Equation (2) holds if and only if Equation (6) holds.

For the imputed reward $\widetilde{R}_t$ defined by Equation (4), Equation (5) implies that it has the same conditional mean reward as R_t given (S_t, A_t):

$$\mathbb{E}[\widetilde{R}_t \mid S_t = s, A_t = a] = \mathbb{E}[b_t(s, a, S_{t+1}) \mid S_t = s, A_t = a] = \mathbb{E}[R_t \mid S_t = s, A_t = a] := \bar{r}_t(s, a).$$

Now consider the augmented process. By Assumption 3.2, the augmented process is Markov, and the Bellman recursion holds with one-step reward $\bar{r}_t(S_t, A_t)$:

$$Q_t^\pi(s, a) = \mathbb{E}\big[\bar{r}_t(s, a) + V_{t+1}^\pi(S_{t+1}, O_t) \mid S_t = s, A_t = a\big],$$
$$V_t^\pi(s, o_-) = \sum_a \pi_t(a \mid s, o_-) Q_t^\pi(s, a), \qquad V_{T+1}^\pi \equiv 0,$$

which is equivalent to Equation (1). Therefore, the policy value $V(\pi) = \mathbb{E}[V_1^\pi(\widetilde{S}_1)]$ is identified.

Part B: Existence of bridge functions

We now establish existence of the bridges. For a probability measure μ, let $\mathcal{L}^2(\mu)$ denote the space of all squared integrable functions of x with respect to measure $\mu(x)$, which is a Hilbert space endowed with the inner product $\langle g_1, g_2 \rangle = \int g_1(x) g_2(x) d\mu(x)$. For any s, a, t, we define operator

$$\mathcal{T}_{t|(s,a)} : \mathcal{L}^2\big(P_{S_{t+1}|s,a}\big) \to \mathcal{L}^2\big(P_{R_t|s,a}\big),$$

where $(\mathcal{T}_{t|(s,a)} h)(r) := \mathbb{E}\big[h(S_{t+1}) \mid R_t = r, S_t = s, A_t = a\big]$. Its adjoint operator is defined by

$$\mathcal{T}^*_{t|(s,a)} : \mathcal{L}^2\big(P_{R_t|s,a}\big) \to \mathcal{L}^2\big(P_{S_{t+1}|s,a}\big),$$

where $(\mathcal{T}^*_{t|(s,a)} g)(s') = \mathbb{E}\big[g(R_t) \mid S_{t+1} = s', S_t = s, A_t = a\big]$. Then, the bridge equation (2) can be written as a first-kind Fredholm integral equation

$$(\mathcal{T}_{t|(s,a)} h)(r) = g(r),$$

where the unknown functions are $h(\cdot) = b_t(s, a, \cdot) \in \mathcal{L}^2\big(P_{S_{t+1}|s,a}\big)$ and the right hand side is $g(r) = r \in \mathcal{L}^2(P_{R_t|s,a})$.

Assumption C.1 (Hilbert-Schmidt property)**.** For any $(s,a) \in \mathcal{S} \times \mathcal{A}$, for all $t = 1, \ldots, T$, denote conditional densities $p_{S_{t+1}|R_t}(s' \mid r, s, a)$, $p_{R_t|S_{t+1}}(r \mid s', s, a)$. We have

$$\int_{\mathcal{R}} \int_{\mathcal{S}} p_{S_{t+1}|R_t}(s' \mid r, s, a) p_{R_t|S_{t+1}}(r \mid s', s, a) ds' dr < \infty.$$

This ensures that the operator $\mathcal{T}_{t|(s,a)}$ is Hilbert–Schmidt and thus admits a singular system $\{(\sigma_{s,a,t,\nu}, \varphi_{s,a,t,\nu}, \psi_{s,a,t,\nu})\}_{\nu \geq 1}$ satisfying $\mathcal{T}_{t|(s,a)} \varphi_{s,a,t,\nu} = \sigma_{s,a,t,\nu} \psi_{s,a,t,\nu}$ and $\mathcal{T}^*_{t|(s,a)} \psi_{s,a,t,\nu} = \sigma_{s,a,t,\nu} \varphi_{s,a,t,\nu}$.

Assumption C.2. Suppose $\{(\sigma_{s,a,t,\nu}, \varphi_{s,a,t,\nu}, \psi_{s,a,t,\nu})\}_{\nu \geq 1}$ is a singular system of $\mathcal{T}_{t|(s,a)}$. Then for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $t = 1, \ldots, T$,

$$\sum_{\nu \geq 1} \frac{\langle g, \psi_{s,a,t,\nu} \rangle^2_{\mathcal{L}^2(P_{R_t|s,a})}}{\sigma^2_{s,a,t,\nu}} < \infty,$$

where $\langle g, \psi_{s,a,t,\nu} \rangle_{\mathcal{L}^2(P_{R_t|s,a})} = \int_{\mathcal{R}} g(r) \cdot \psi_{s,a,t,\nu}(r) p_{R_t|s,a}(r) dr$.

Lemma C.3 (Picard's Theorem (Kress, 1989))**.** *Let $\mathcal{H}_1, \mathcal{H}_2$ be real Hilbert spaces and $K : \mathcal{H}_1 \to \mathcal{H}_2$ a compact linear operator with adjoint $K^* : \mathcal{H}_2 \to \mathcal{H}_1$. Then, there exists a singular system $\{(\lambda_\nu, \phi_\nu, \psi_\nu)\}_{\nu=1}^\infty$ of K, with singular values $\lambda_\nu > 0$ and orthonormal sequences $\{\phi_\nu\} \subset \mathcal{H}_1$, $\{\psi_\nu\} \subset \mathcal{H}_2$ satisfying*

$$K\phi_\nu = \lambda_\nu \psi_\nu, \quad K^* \psi_\nu = \lambda_\nu \phi_\nu.$$

Given $g \in \mathcal{H}_2$, the first–kind Fredholm equation $Kh = g$ has a solution $h \in \mathcal{H}_1$ if and only if

(a) $g \in \ker(K^*)^\perp$;

(b) $\sum_{\nu=1}^\infty \lambda_\nu^{-2} |\langle g, \psi_\nu \rangle_{\mathcal{H}_2}|^2 < \infty$,

where $\ker(K^*) = \{h : K^* h = 0\}$ is the null space of K^*, and $\perp$ denotes the orthogonal complement to a set.

Proposition C.4 (Existence of bridges)**.** *Under Assumption 4.4 (1), Assumptions C.1 and C.2, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $t = 1, \ldots, T$, there exists a solution h to equation*

$$\mathcal{T}_{t|(s,a)} h = g,$$

where $h := b_t(s, a; \cdot) \in \mathcal{L}^2\big(P_{S_{t+1}|s,a}\big)$ and $g(r) = r \in \mathcal{L}^2(P_{R_t|s,a})$. Equivalently, for any fixed (s,a) and t, there exists a function $b_t(s, a, \cdot)$ satisfying Equation (2).

Proof. By Assumption C.1, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $t = 1, \ldots, T$, the operator $\mathcal{T}_{t|(s,a)}$ is Hilbert–Schmidt and thus compact. Suppose there exists function $f \in \ker(\mathcal{T}^*_{t|(s,a)})$, then by definition, $\mathcal{T}^*_{t|(s,a)} f = 0$. By Assumption 4.4 (1), we have $f(R_t) = 0, \ a.s.$ Therefore, $\ker(\mathcal{T}^*_{t|(s,a)}) = \{0\}$, and hence $\ker(\mathcal{T}^*_{t|(s,a)})^\perp = \mathcal{L}^2\big(P_{R_t|s,a}\big)$. Because reward R_t is bounded, then $g \in \mathcal{L}^2\big(P_{R_t|s,a}\big)$. So the condition (a) in Lemma C.3 is satisfied.

Additionally, condition (b) is also satisfied by Assumption C.2. By Lemma C.3, there exists a solution $h \in \mathcal{L}^2(P_{S_{t+1}|s,a})$ to $\mathcal{T}_{t|(s,a)} h = g$ where $g(r) = r$, i.e., there exists function $b_t(s, a, \cdot)$ such that:

$$\mathbb{E}\big[b_t(s, a, S_{t+1}) \mid R_t = r, S_t = s, A_t = a\big] = r.$$

□

Part C: Uniqueness of bridge functions

Next we study the uniqueness of the bridge functions.

Proposition C.5 (Uniqueness of bridges)**.** *Under Assumption 4.4 (2), any bridge function b_t satisfying*

$$\mathbb{E}[b_t(S_t, A_t, S_{t+1}) \mid R_t, S_t, A_t] = R_t, \quad a.s.$$

is unique.

Proof. Suppose there exist different bridge functions $b_{1,t}$ and $b_{2,t}$ satisfying the equation above for any t. Then,

$$\mathbb{E}[b_{1,t}(S_t, A_t, S_{t+1}) - b_{2,t}(S_t, A_t, S_{t+1}) \mid R_t, S_t, A_t] = 0, \quad a.s.$$

By Assumption 4.4 (2), we have

$$b_{1,t} - b_{2,t} = 0, \quad a.s., \ \ \forall t = 1, \dots, T,$$

which contradicts $b_{1,t} \neq b_{2,t}$. Thus uniqueness holds. □

D. Proof of Theorem 6.5

For a function class $\mathcal{G}$ and radius $\delta > 0$, given sample $\{X_i\}$, the *local empirical Rademacher complexity* is defined by

$$\widehat{\mathcal{R}}_n(\mathcal{G}, \delta) = \mathbb{E}_\varepsilon\Big[\sup_{g \in \mathcal{G}: \|g\|_{2,n} \leq \delta} \big|\frac{1}{n}\sum_{i=1}^n \varepsilon_i g(X_i)\big| \,\Big|\, \{X_i\}\Big]$$

where $\|g\|_{2,n}^2 = \frac{1}{n}\sum_{i=1}^n g(X_i)^2$. The empirical critical radius $\hat{\delta}_n$ is the smallest solution to the inequality $\widehat{\mathcal{R}}_n(\mathcal{G}, \delta) \leq \frac{\delta^2}{b}$. Wainwright (2019) gives the relationship of critical radius and empirical critical radius: with probability at least $1 - \zeta$,

$$\delta_n \leq \mathcal{O}(\hat{\delta}_n + \sqrt{\frac{\log(1/\zeta)}{n}}),$$

which enables us to study on empirical critical radius $\hat{\delta}_n$.

Define

$$\Psi_n^t(b, g) = \frac{1}{n_t}\sum_{i \in \mathcal{I}_t^{\text{obs}}} (b_t(S_{t,i}, A_{t,i}, S_{t+1,i}) - R_{t,i}^{\text{obs}}) g_t(R_{t,i}^{\text{obs}}, S_{t,i}, A_{t,i}),$$

and population level version

$$\Psi^t(b, g) = \mathbb{E}\Big[(b_t(S_t, A_t, S_{t+1}) - R_t) g_t(R_t, S_t, A_t) \,\Big|\, O_t = 1\Big].$$

Moreover, Let

$$\Psi_n^{t,\lambda}(b, g) = \Psi_n^t(b, g) - \lambda\left(\|g_t\|_{\mathcal{G}^{(t)}}^2 + \frac{U}{\delta^2}\|g_t\|_{2,n_t}^2\right),$$

$$\Psi^{t,\lambda}(b, g) = \Psi^t(b, g) - \lambda\left(\frac{2}{3}\|g_t\|_{\mathcal{G}^{(t)}}^2 + \frac{U}{2\delta^2}\|g_t\|_2^2\right).$$

So the minimizer $\hat{b}_t$ can be written as

$$\hat{b}_t = \arg\min_{b_t \in \mathcal{B}^{(t)}} \sup_{g_t \in \mathcal{G}^{(t)}} \Psi_n^{t,\lambda}(b, g) + \lambda\mu\|b_t\|_{\mathcal{B}^{(t)}}^2.$$

By Lemma G.2, with probability at least $1 - \zeta$, for any function $g_t \in \mathcal{G}_{3U}^{(t)}$,

$$\big|\|g_t\|_{2,n_t}^2 - \|g_t\|_2^2\big| \leq \frac{1}{2}\|g_t\|_2^2 + (\delta_t^{\mathcal{G}})^2,$$

where $\delta_t^{\mathcal{G}} = \delta_{n_t}^{\mathcal{G}} + c_0\sqrt{\frac{\log(c_1/\zeta)}{n_t}}$, and $\delta_{n_t}^{\mathcal{G}}$ is the upper bound of the empirical critical radii of function class $\mathcal{G}_{3U}^{(t)}$. For any $\|g_t\|_{\mathcal{G}^{(t)}}^2 \geq 3U$, consider rescaling g_t by $\frac{\sqrt{3U}}{\|g_t\|_{\mathcal{G}^{(t)}}} g_t \in \mathcal{G}_{3U}^{(t)}$, and we have

$$\big|\|g_t\|_{2,n_t}^2 - \|g_t\|_2^2\big| \leq \frac{1}{2}\|g_t\|_2^2 + (\delta_t^{\mathcal{G}})^2\frac{\|g_t\|_{\mathcal{G}^{(t)}}^2}{3U}.$$

Combine the above inequalities,

$$\big|\|g_t\|_{2,n_t}^2 - \|g_t\|_2^2\big| \leq \frac{1}{2}\|g_t\|_2^2 + (\delta_t^{\mathcal{G}})^2 \max\left\{1, \frac{\|g_t\|_{\mathcal{G}^{(t)}}^2}{3U}\right\}. \tag{9}$$

Thus,

$$\begin{aligned}\|g_t\|^2_{\mathcal{G}^{(t)}} + \frac{U}{\delta^2}\|g_t\|^2_{2,n_t} &\geq \|g_t\|^2_{\mathcal{G}^{(t)}} + \frac{U}{(\delta_t^{\mathcal{G}})^2}\left[\frac{1}{2}\|g_t\|_2^2 - (\delta_t^{\mathcal{G}})^2 \max\left\{1, \frac{\|g_t\|^2_{\mathcal{G}^{(t)}}}{3U}\right\}\right] \\ &\geq \frac{2}{3}\|g_t\|^2_{\mathcal{G}^{(t)}} + \frac{U}{2(\delta_t^{\mathcal{G}})^2}\|g_t\|_2^2 - U.\end{aligned} \tag{10}$$

Next, we study the upper and lower bounds of the centered empirical sup-loss

$$\sup_{g_t \in \mathcal{G}^{(t)}} \Psi_n^t(\hat{b}_t, g) - \Psi_n^t(b_t^*, g) - 2\lambda\left(\|g_t\|^2_{\mathcal{G}^{(t)}} + \frac{U}{(\delta_t^{\mathcal{G}})^2}\|g_t\|^2_{2,n_t}\right).$$

For simplicity we omit t and write it as

$$\sup_{g \in \mathcal{G}} \Psi_n(\hat{b}, g) - \Psi_n(b^*, g) - 2\lambda\left(\|g\|^2_{\mathcal{G}} + \frac{U}{\delta^2}\|g\|^2_{2,n}\right). \tag{11}$$

D.1. Upper bounding the centered empirical sup-loss

We first decompose $\Psi_n^\lambda(b, g)$ by

$$\begin{aligned}\Psi_n^\lambda(b, g) &= \Psi_n(b, g) - \Psi_n(b^*, g) + \Psi_n(b^*, g) - \lambda\left(\|g\|^2_{\mathcal{G}} + \frac{U}{\delta^2}\|g\|^2_{2,n}\right) \\ &\geq \Psi_n(b, g) - \Psi_n(b^*, g) - 2\lambda\left(\|g\|^2_{\mathcal{G}} + \frac{U}{\delta^2}\|g\|^2_{2,n}\right) - \sup_{g \in \mathcal{G}} \Psi_n^\lambda(b^*, g),\end{aligned}$$

where the last inequality holds by symmetry of $\mathcal{G}$. Then we have

$$\begin{aligned}\sup_{g \in \mathcal{G}} \Psi_n(\hat{b}, g) - \Psi_n(b^*, g) - 2\lambda\left(\|g\|^2_{\mathcal{G}} + \frac{U}{\delta^2}\|g\|^2_{2,n}\right) &\leq \sup_{g \in \mathcal{G}} \Psi_n^\lambda(b^*, g) + \Psi_n^\lambda(\hat{b}, g) \\ &\leq \sup_{g \in \mathcal{G}} \Psi_n^\lambda(b^*, g) + \big[\sup_{g \in \mathcal{G}} \Psi_n^\lambda(b^*, g) + \lambda\mu\big(\|b^*\|^2_{\mathcal{B}} - \|\hat{b}\|^2_{\mathcal{B}}\big)\big] \\ &\leq 2\sup_{g \in \mathcal{G}} \Psi_n^\lambda(b^*, g) + \lambda\mu\big(\|b^*\|^2_{\mathcal{B}} - \|\hat{b}\|^2_{\mathcal{B}}\big).\end{aligned} \tag{12}$$

By Lemma G.3, for all $g \in \mathcal{G}$, similarly, with probability at least $1 - \zeta$ we have

$$\begin{aligned}\big|\Psi_n(b^*, g) - \Psi(b^*, g)\big| &\leq 36\delta\big[\|g\|_2 + \delta \max\left\{1, \frac{\|g\|_{\mathcal{G}}}{\sqrt{3U}}\right\}\big] \\ &\leq 36\delta\big[\|g\|_2 + \delta(1 + \frac{\|g\|_{\mathcal{G}}}{\sqrt{3U}})\big].\end{aligned} \tag{13}$$

Combine Equation (10) and Equation (13), we have with probability at least $1 - 2\zeta$, for all $b \in \mathcal{B}$ and $g \in \mathcal{G}$,

$$\begin{aligned}\Psi_n^\lambda(b^*, g) &= \Psi_n(b^*, g) - \lambda\left(\|g\|^2_{\mathcal{G}} + \frac{U}{\delta^2}\|g\|^2_{2,n}\right) \\ &\leq \Psi(b^*, g) + |\Psi_n(b^*, g) - \Psi(b^*, g)| - \lambda\left(\|g\|^2_{\mathcal{G}} + \frac{U}{\delta^2}\|g\|^2_{2,n}\right) \\ &\leq \Psi(b^*, g) + 36\delta\big[\|g\|_2 + \delta(1 + \frac{\|g\|_{\mathcal{G}}}{\sqrt{3U}})\big] - \lambda\left(\|g\|^2_{\mathcal{G}} + \frac{U}{\delta^2}\|g\|^2_{2,n}\right) \\ &\leq \Psi(b^*, g) + 36\delta\big[\|g\|_2 + \delta(1 + \frac{\|g\|_{\mathcal{G}}}{\sqrt{3U}})\big] - \lambda\left(\frac{2}{3}\|g\|^2_{\mathcal{G}} + \frac{U}{2\delta^2}\|g\|^2_{2,n}\right) + \lambda U \\ &\leq \Psi(b^*, g) - \lambda\left(\frac{1}{3}\|g\|^2_{\mathcal{G}} + \frac{U}{4\delta^2}\|g\|^2_{2,n}\right) + 36\delta^2 + \lambda U + 36\delta\|g\|_2 + 36\delta^2\frac{\|g\|_{\mathcal{G}}}{\sqrt{3U}} - \lambda\frac{U}{4\delta^2}\|g\|^2_{2,n} - \frac{\lambda}{3}\|g\|^2_{\mathcal{G}} \\ &= \Psi^{\lambda/2}(b^*, g) + 36\delta^2 + \lambda U + \big(36\delta\|g\|_2 - \lambda\frac{U}{4\delta^2}\|g\|^2_{2,n}\big) + \big(36\delta^2\frac{\|g\|_{\mathcal{G}}}{\sqrt{3U}} - \frac{\lambda}{3}\|g\|^2_{\mathcal{G}}\big).\end{aligned}$$

Using the fact that for any $a, b > 0$ and any norm, $\sup_{g\in\mathcal{G}}(a\|g\| - b\|g\|^2) \le \frac{a^2}{4b}$, suppose $\lambda \ge \frac{C_1\delta^2}{U}$, and then

$$\sup_{g\in\mathcal{G}}\big(36\delta\|g\|_2 - \lambda\frac{U}{4\delta^2}\|g\|_{2,n}^2\big) \le \frac{36^2\delta^4}{\lambda U} \le \frac{36^2\delta^2}{C_1};$$
$$\sup_{g\in\mathcal{G}}\big(36\delta^2\frac{\|g\|_{\mathcal{G}}}{\sqrt{3U}} - \frac{\lambda}{3}\|g\|_{\mathcal{G}}^2\big) \le \frac{18^2\delta^4}{\lambda U} \le \frac{18^2\delta^2}{C_1}.$$

Therefore,

$$\Psi_n^{\lambda}(b^*, g) \le \Psi^{\lambda/2}(b^*, g) + 36\delta^2 + \lambda U + \frac{5\times 18^2\delta^2}{C_1}$$

Combine this with Equation (12) we get

$$\begin{aligned}\sup_{g\in\mathcal{G}}\Psi_n(\hat{b}, g) - \Psi_n(b^*, g) - 2\lambda\left(\|g\|_{\mathcal{G}}^2 + \frac{U}{\delta^2}\|g\|_{2,n}^2\right) &\le 2\sup_{g\in\mathcal{G}}\Psi_n^{\lambda}(b^*, g) + \lambda\mu\big(\|b^*\|_{\mathcal{B}}^2 - \|\hat{b}\|_{\mathcal{B}}^2\big)\\ &\le 2\sup_{g\in\mathcal{G}}\Psi^{\lambda/2}(b^*, g) + 2\lambda U + (72 + \frac{10\times 18^2}{C_1})\delta^2 + \lambda\mu\big(\|b^*\|_{\mathcal{B}}^2 - \|\hat{b}\|_{\mathcal{B}}^2\big)\\ &= 2\lambda U + (72 + \frac{10\times 18^2}{C_1})\delta^2 + \lambda\mu\big(\|b^*\|_{\mathcal{B}}^2 - \|\hat{b}\|_{\mathcal{B}}^2\big).\end{aligned} \tag{14}$$

D.2. Lower bounding the centered empirical sup-loss

By Assumption 6.4, we write $g_b = \arg\inf_{g\in\mathcal{G}_{L^2\|b-b^*\|_{\mathcal{B}}^2}}\|g - \mathcal{T}(b-b^*)\|_2$ where

$$\sup_{b\in\mathcal{B}}\inf_{g\in\mathcal{G}_{L^2\|b-b^*\|_{\mathcal{B}}^2}}\|g - \mathcal{T}(b-b^*)\|_2 \le \eta < \infty.$$

Also, let $g_{\hat{b}} = \arg\inf_{g\in\mathcal{G}_{L^2\|\hat{b}-b^*\|_{\mathcal{B}}^2}}\|g - \mathcal{T}(\hat{b}-b^*)\|_2$.

When $\|g_{\hat{b}}\|_2 \le \delta$, then

$$\|\mathcal{T}(\hat{b}-b^*)\|_2 \le \|g_{\hat{b}}\| + \|g - \mathcal{T}(b-b^*)\|_2 \le \delta + \eta;$$

if $\|g_{\hat{b}}\|_2 \ge \delta$, let $r = \frac{\delta}{2\|g_{\hat{b}}\|_2} \in [0, \frac{1}{2}]$, and $rg_{\hat{b}} \in \mathcal{G}_{L^2\|\hat{b}-b^*\|_{\mathcal{B}}^2}$ since $\mathcal{G}$ is star-shaped. Hence,

$$\begin{aligned}\sup_{g\in\mathcal{G}}\Psi_n(\hat{b}, g) - \Psi_n(b^*, g) - 2\lambda\left(\|g\|_{\mathcal{G}}^2 + \frac{U}{\delta^2}\|g\|_{2,n}^2\right) &\ge \Psi_n(\hat{b}, rg_{\hat{b}}) - \Psi_n(b^*, rg_{\hat{b}}) - 2\lambda\left(\|rg_{\hat{b}}\|_{\mathcal{G}}^2 + \frac{U}{\delta^2}\|rg_{\hat{b}}\|_{2,n}^2\right)\\ &= \underbrace{r\left(\Psi_n(\hat{b}, g_{\hat{b}}) - \Psi_n(b^*, g_{\hat{b}})\right)}_{(i)} - 2\lambda\underbrace{r^2\left(\|g_{\hat{b}}\|_{\mathcal{G}}^2 + \frac{U}{\delta^2}\|g_{\hat{b}}\|_{2,n}^2\right)}_{(ii)}.\end{aligned}$$

For (ii),

$$r^2\left(\|g_{\hat{b}}\|_{\mathcal{G}}^2 + \frac{U}{\delta^2}\|g_{\hat{b}}\|_{2,n}^2\right) \le \frac{1}{4}\|g_{\hat{b}}\|_{\mathcal{G}}^2 + \frac{r^2U}{\delta^2}\|g_{\hat{b}}\|_{2,n}^2. \tag{15}$$

By Equation (9), with probability at least $1-\zeta$,

$$\frac{r^2U}{\delta^2}\|g_{\hat{b}}\|_{2,n}^2 \le \frac{r^2U}{\delta^2}\Big[\|g_{\hat{b}}\|_2^2 + \big|\|g_{\hat{b}}\|_{2,n}^2 - \|g_{\hat{b}}\|_2^2\big|\Big] \le \frac{r^2U}{\delta^2}\big(\frac{3}{2}\|g_{\hat{b}}\|_2^2 + \delta^2(1 + \frac{\|g_{\hat{b}}\|_{\mathcal{G}}^2}{3U})\big).$$

Substitute this into Equation (15) and we get

$$
\begin{aligned}
r^2\left(\|g_{\hat{b}}\|_{\mathcal{G}}^2+\frac{U}{\delta^2}\|g_{\hat{b}}\|_{2,n}^2\right) &\le \frac{1}{4}\|g_{\hat{b}}\|_{\mathcal{G}}^2+\frac{r^2U}{\delta^2}\left(\frac{3}{2}\|g_{\hat{b}}\|_2^2+\delta^2\left(1+\frac{\|g_{\hat{b}}\|_{\mathcal{G}}^2}{3U}\right)\right)\\
&\le \left(\frac{1}{4}\|g_{\hat{b}}\|_{\mathcal{G}}^2+\frac{1}{12}\|g_{\hat{b}}\|_{\mathcal{G}}^2\right)+\frac{3Ur^2}{2\delta^2}\|g_{\hat{b}}\|_2^2+\frac{1}{4}U\\
&= \frac{1}{3}\|g_{\hat{b}}\|_{\mathcal{G}}^2+\left(\frac{3}{8}+\frac{1}{4}\right)U\\
&\le \frac{1}{3}L^2\|\hat{b}-b^*\|_{\mathcal{B}}^2+\frac{5}{8}U.
\end{aligned}
\tag{16}
$$

For (i), we consider function class

$$
\mathcal{J}_{B,L^2B}=\left\{((s,a,s'),(r,s,a))\mapsto\alpha(b(s,a,s')-b^*(s,a,s'))g_b^{L^2B}(r,s,a)\mid b-b^*\in\mathcal{B}_B,\alpha\in[0,1]\right\},
$$

where $g_b^{L^2B}(r,s,a)=\arg\inf_{g\in\mathcal{G}_{L^2B}}\|g-\mathcal{T}(b-b^*)\|_2$. Choose $\delta=\delta_n^{\mathcal{J}}+c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$, where $\delta_n^{\mathcal{J}}$ is the upper bound of the empirical critical radii of function class $\mathcal{J}_{B,L^2B}$. Choose loss function $\mathcal{L}=(b-b^*)f$, then by Lemma G.3, we have with probability at least $1-\zeta$, for all $b\in\mathcal{B}$ and $g\in\mathcal{G}$,

$$
\begin{aligned}
|(\Psi_n(b,g_b)-\Psi_n(b^*,g_b))-(\Psi(b,g_b)-\Psi(b^*,g_b))| &\le 18\delta(\|(b^*-b)g_b\|_2+\delta)\\
&\le 18\delta(\|g_b\|_2+\delta),
\end{aligned}
$$

since $b-b^*\in\mathcal{B}_B$, which is 1-uniformly bounded.

When $\|b-b^*\|_{\mathcal{B}}^2>B$, we rescale the function by $\frac{\sqrt{B}}{\|b-b^*\|_{\mathcal{B}}}(b-b^*)$, and similarly, we obtain that with probability at least $1-\zeta$,

$$
|(\Psi_n(b,g_b)-\Psi_n(b^*,g_b))-(\Psi(b,g_b)-\Psi(b^*,g_b))|\le 18\delta(\|g_b\|_2+\delta)\max\left\{1,\frac{\|b-b^*\|_{\mathcal{B}}^2}{B}\right\}.
$$

Therefore, with probability at least $1-\zeta$, for any $g\in\mathcal{G}$,

$$
\begin{aligned}
r\left(\Psi_n(\hat{b},g_{\hat{b}})-\Psi_n(b^*,g_{\hat{b}})\right) &\ge r\left(\Psi(\hat{b},g_{\hat{b}})-\Psi(b^*,g_{\hat{b}})\right)-r|(\Psi_n(b,g_{\hat{b}})-\Psi_n(b^*,g_{\hat{b}}))-(\Psi(b,g_{\hat{b}})-\Psi(b^*,g_{\hat{b}}))|\\
&\ge \underbrace{r\left(\Psi(\hat{b},g_{\hat{b}})-\Psi(b^*,g_{\hat{b}})\right)}_{(A)}-\underbrace{18\delta r(\|g_{\hat{b}}\|_2+\delta)\max\left\{1,\frac{\|\hat{b}-b^*\|_{\mathcal{B}}^2}{B}\right\}}_{(B)}.
\end{aligned}
$$

For (A),

$$
\begin{aligned}
r\left(\Psi(\hat{b}, g_{\hat{b}}) - \Psi(b^*, g_{\hat{b}})\right) &= \frac{\delta}{2\|g_{\hat{b}}\|_2}\mathbb{E}\Big[(\hat{b}(S,A,S') - b^*(S,A,S'))g_{\hat{b}}(R,S,A)\,\Big|\,O_t = 1\Big] \\
&= \frac{\delta}{2\|g_{\hat{b}}\|_2}\mathbb{E}\Big[g_{\hat{b}}(R,S,A)\mathbb{E}\left(\hat{b}(S,A,S') - b^*(S,A,S') \mid R,S,A,O_t = 1\right)\,\Big|\,O_t = 1\Big] \\
&= \frac{\delta}{2\|g_{\hat{b}}\|_2}\mathbb{E}\Big\{g_{\hat{b}}(R,S,A)\big[\mathcal{T}(\hat{b} - b^*)(R,S,A)\big]\Big\} \\
&= \frac{\delta}{2\|g_{\hat{b}}\|_2}\mathbb{E}\big\{(g_{\hat{b}}(R,S,A))^2 - g_{\hat{b}}(R,S,A)[g_{\hat{b}}(R,S,A) - \mathcal{T}(\hat{b} - b^*)(R,S,A)]\big\} \\
&= \frac{\delta}{2\|g_{\hat{b}}\|_2}\Big\{\|g_{\hat{b}}\|_2^2 - \big\{\mathbb{E}g_{\hat{b}}(R,S,A)[g_{\hat{b}}(R,S,A) - \mathcal{T}(\hat{b} - b^*)(R,S,A)]\big\}\Big\} \\
&\geq \frac{\delta}{2\|g_{\hat{b}}\|_2}\Big\{\|g_{\hat{b}}\|_2^2 - \|g_{\hat{b}}\|_2\|g_{\hat{b}} - \mathcal{T}(\hat{b} - b^*))\|_2\Big\} \\
&= \frac{\delta}{2}\Big\{\|g_{\hat{b}}\|_2 - \|g_{\hat{b}} - \mathcal{T}(\hat{b} - b^*))\|_2\Big\} \\
&\geq \frac{\delta}{2}\Big\{\|\mathcal{T}(\hat{b} - b^*)\|_2 - 2\|g_{\hat{b}} - \mathcal{T}(\hat{b} - b^*))\|_2\Big\} \\
&\geq \frac{\delta}{2}\Big\{\|\mathcal{T}(\hat{b} - b^*)\|_2 - 2\eta\Big\}.
\end{aligned}
$$

For (B),

$$
\begin{aligned}
18\delta r(\|g_{\hat{b}}\|_2 + \delta)\max\Big\{1, \frac{\|\hat{b} - b^*\|_{\mathcal{B}}^2}{B}\Big\} &= 18\delta r(\frac{\delta}{2r} + \delta)\max\Big\{1, \frac{\|\hat{b} - b^*\|_{\mathcal{B}}^2}{B}\Big\} \\
&= (9\delta^2 + 18r\delta^2)\max\Big\{1, \frac{\|\hat{b} - b^*\|_{\mathcal{B}}^2}{B}\Big\} \\
&\leq 18\delta^2 + \frac{18\delta^2\|\hat{b} - b^*\|_{\mathcal{B}}^2}{B}.
\end{aligned}
$$

So when $\|g_{\hat{b}}\|_2 \geq \delta$, with probability at least $1 - 2\zeta$,

$$
\begin{aligned}
(i) = r\left(\Psi_n(\hat{b}, g_{\hat{b}}) - \Psi_n(b^*, g_{\hat{b}})\right) &\geq (A) - (B) \\
&\geq \frac{\delta}{2}\Big\{\|\mathcal{T}(\hat{b} - b^*)\|_2 - 2\eta\Big\} - 18\delta^2 - \frac{18\delta^2\|\hat{b} - b^*\|_{\mathcal{B}}^2}{B}.
\end{aligned}
$$

Therefore the lower bound of $\sup_{g\in\mathcal{G}} \Psi_n(\hat{b}, g) - \Psi_n(b^*, g) - 2\lambda\left(\|g\|_{\mathcal{G}}^2 + \frac{U}{\delta^2}\|g\|_{2,n}^2\right)$ is given by

$$
\begin{aligned}
&\sup_{g\in\mathcal{G}} \Psi_n(\hat{b}, g) - \Psi_n(b^*, g) - 2\lambda\left(\|g\|_{\mathcal{G}}^2 + \frac{U}{\delta^2}\|g\|_{2,n}^2\right) \\
\geq &(i) - 2\lambda(ii) \\
\geq &\frac{\delta}{2}\Big\{\|\mathcal{T}(\hat{b} - b^*)\|_2 - 2\eta\Big\} - 18\delta^2 - \frac{18\delta^2\|\hat{b} - b^*\|_{\mathcal{B}}^2}{B} - 2\lambda\big(\frac{1}{3}L^2\|\hat{b} - b^*\|_{\mathcal{B}}^2 + \frac{5}{8}U\big) \\
\geq &\frac{\delta}{2}\|\mathcal{T}(\hat{b} - b^*)\|_2 - \eta\delta - 18\delta^2 - (\frac{18\delta^2}{B} + \frac{2\lambda L^2}{3})\|\hat{b} - b^*\|_{\mathcal{B}}^2 - \frac{5}{4}\lambda U.
\end{aligned}
$$

D.3. Combining the upper and lower bounds

Combine the upper bound and lower bound, and then we have either $\|g_{\hat{b}}\|_2 \le \delta$, or with probability at least $1-3\zeta$, for all $b \in \mathcal{B}$,

$$\frac{\delta}{2}\|\mathcal{T}(\hat{b}-b^*)\|_2 - \eta\delta - 18\delta^2 - (\frac{18\delta^2}{B} + \frac{2\lambda L^2}{3})\|\hat{b}-b^*\|_{\mathcal{B}}^2 - \frac{5}{4}\lambda U \le 2\lambda U + (72 + \frac{10\times 18^2}{C_1})\delta^2 + \lambda\mu\big(\|b^*\|_{\mathcal{B}}^2 - \|\hat{b}\|_{\mathcal{B}}^2\big).$$

So

$$\begin{aligned}\frac{\delta}{2}\|\mathcal{T}(\hat{b}-b^*)\|_2 &\le \frac{13}{4}\lambda U + \eta\delta + \lambda\mu\big(\|b^*\|_{\mathcal{B}}^2 - \|\hat{b}\|_{\mathcal{B}}^2\big) + (90 + \frac{10\times 18^2}{C_1})\delta^2 + (\frac{18\delta^2}{B} + \frac{2\lambda L^2}{3})\|\hat{b}-b^*\|_{\mathcal{B}}^2 \\ &\le \frac{13}{4}\lambda U + \eta\delta + \lambda\mu\big(\|b^*\|_{\mathcal{B}}^2 - \|\hat{b}\|_{\mathcal{B}}^2\big) + (90 + \frac{10\times 18^2}{C_1})\delta^2 + 2\lambda(\frac{18\delta^2}{\lambda B} + \frac{2L^2}{3})\big(\|b^*\|_{\mathcal{B}}^2 + \|\hat{b}\|_{\mathcal{B}}^2\big).\end{aligned}$$

If $\mu \ge \frac{36\delta^2}{\lambda B} + \frac{4L^2}{3}$, then

$$\begin{aligned}\|\mathcal{T}(\hat{b}-b^*)\|_2 &\le \frac{2}{\delta}\Big(\frac{13}{4}\lambda U + \eta\delta + \lambda\mu\big(\|b^*\|_{\mathcal{B}}^2 - \|\hat{b}\|_{\mathcal{B}}^2\big) + (90 + \frac{10\times 18^2}{C_1})\delta^2 + \lambda\mu\big(\|b^*\|_{\mathcal{B}}^2 + \|\hat{b}\|_{\mathcal{B}}^2\big)\Big) \\ &= \frac{2}{\delta}\Big(\frac{13}{4}\lambda U + \eta\delta + 2\lambda\mu\|b^*\|_{\mathcal{B}}^2 + (90 + \frac{10\times 18^2}{C_1})\delta^2\Big) \\ &\le \frac{13}{2}\frac{\lambda U}{\delta} + 2\eta + 4\frac{\lambda\mu}{\delta}\|b^*\|_{\mathcal{B}}^2 + (180 + \frac{20\times 18^2}{C_1})\delta.\end{aligned}$$

Suppose $\lambda \le \frac{C_2\delta^2}{U}$, then with probability at least $1-4\zeta$,

$$\begin{aligned}\|\mathcal{T}(\hat{b}-b^*)\|_2 &\le \frac{13}{2}\frac{\lambda U}{\delta} + 2\eta + 4\frac{\lambda\mu}{\delta}\|b^*\|_{\mathcal{B}}^2 + (180 + \frac{20\times 18^2}{C_1})\delta \\ &\le \frac{13}{2}C_2\delta + 2\eta + 4C_2\mu\delta\|b^*\|_{\mathcal{B}}^2 + (180 + \frac{20\times 18^2}{C_1})\delta \\ &\le 4C_2\mu\delta\|b^*\|_{\mathcal{B}}^2 + (\frac{13}{2}C_2 + 180 + \frac{20\times 18^2}{C_1})\delta + 2\eta \\ &\lesssim \delta\max\Big\{1, \|b^*\|_{\mathcal{B}}^2\Big\}.\end{aligned}$$

E. Proof of Corollary 6.6

Recall that the product class is denoted by

$$\mathcal{J}_{B,L^2B}^{(t)} := \Big\{((s,a,s'),(r,s,a)) \mapsto \alpha(b_t(s,a,s') - b_t^*(s,a,s'))g_{b,t}^{L^2B}(r,s,a) \mid b_t - b_t^* \in \mathcal{B}_B^{(t)}, \alpha \in [0,1]\Big\}.$$

Define the tensor product of two RKHSs $\mathcal{B}^{(t)}$ and $\mathcal{G}^{(t)}$ as $\mathcal{H}_{\otimes}^{(t)}$ endowed with kernel $K_{\otimes,t}\big((x,y),(x',y')\big) := K_{B,t}(x,x')\,K_{G,t}(y,y')$. Then, one can verify that $\mathcal{J}_{B,L^2B}^{(t)}$ satisfies

$$\mathcal{J}_{B,L^2B}^{(t)} \subseteq \Big\{f \in \mathcal{H}_{\otimes}^{(t)} :\ \|f\|_{\mathcal{H}_{\otimes}^{(t)}} \le \sqrt{B_t}\sqrt{L^2B_t}\Big\} =: \mathcal{H}_{\otimes,LB}^{(t)}.$$

By Lemma G.5, we have that

$$\mathcal{R}_{n_t}(\mathcal{J}_{B,L^2B}^{(t)},\delta) \le LB\sqrt{\frac{2}{n_t}}\sqrt{\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\min\big\{\lambda_{t,i}^{\mathcal{B}}\lambda_{t,j}^{\mathcal{G}},\delta^2\big\}}. \tag{17}$$

Under the polynomial eigen decay assumptions $\lambda_{t,i}^{B} \lesssim i^{-2\alpha_B}$ and $\lambda_{t,j}^{G} \lesssim j^{-2\alpha_G}$ with $\alpha_{\min} := \min\{\alpha_B,\alpha_G\}$, by Lemma G.6, the tensor-product spectrum admits the bound

$$\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\min\{\lambda_{t,i}^{B}\lambda_{t,j}^{G},\delta^2\} \lesssim \delta^{2-\frac{1}{\alpha_{\min}}}\log\frac{1}{\delta}.$$

plugging this into Equation (17) yields

$$\delta_{n_t}^{\mathcal{J}} \lesssim LB\, n_t^{-\frac{\alpha_{\min}}{2\alpha_{\min}+1}} \log n_t.$$

Similarly, we have

$$\delta_{n_t}^{\mathcal{G}} \lesssim \sqrt{U_t}\, n_t^{-\frac{\alpha_G}{2\alpha_G+1}} \log n_t.$$

Therefore, $\delta_{n_t} = \max\{\delta_{n_t}^{\mathcal{G}}, \delta_{n_t}^{\mathcal{J}}\}$ satisfies

$$\delta_{n_t} \lesssim \max\{\sqrt{U_t}, LB\}\, n_t^{-\frac{\alpha_{\min}}{2\alpha_{\min}+1}} \log n_t,$$

and the claim follows by Theorem 6.5.

Note that the min-max estimation problem of b_t has a closed-form solution, which is discussed in Dikkala et al. (2020) Appendix E.3.

F. Proof of Theorem 6.9

F.1. Error Decomposition

The estimation error bound can be decomposed as

$$|\mathbb{E}[V_1^\pi] - \widehat{V}(\pi)| \le \underbrace{|\mathbb{E}[V_1^\pi] - \mathbb{E}_n[V_1^\pi]|}_{(I)} + \underbrace{|\mathbb{E}[V_1^\pi] - \mathbb{E}[\widehat{V}_1^\pi]|}_{(II)} + \underbrace{|\mathbb{E}(V_1^\pi - \widehat{V}_1^\pi) - \mathbb{E}_n(V_1^\pi - \widehat{V}_1^\pi)|}_{(III)},$$

where $\mathbb{E}_n[V_1^\pi] := \frac{1}{n}\sum_{i=1}^n V_1^\pi(S_{1,i}, 0)$.

F.2. Bound of (I)

For (I), since rewards are bounded in $[-1, 1]$, by Hoeffding inequality, with probability at least $1-\zeta$,

$$|\mathbb{E}[V_1^\pi] - \mathbb{E}_n[V_1^\pi]| \lesssim \|V_1^\pi\|_\infty \sqrt{\frac{\log(c_1/\zeta)}{n}} \le T\sqrt{\frac{\log(c_1/\zeta)}{n}},$$

where constant $c_1 > 0$.

F.3. Bound of (II)

For (II),

$$\begin{aligned}
|\mathbb{E}[V_1^\pi] - \mathbb{E}[\widehat{V}_1^\pi]| &\le \|V_1^\pi - \widehat{V}_1^\pi\|_2 \\
&= \|\sum_a \pi_1(a \mid S_1, O_0)(Q_1^\pi(S_1, a) - \widehat{Q}_1(S_1, a))\|_2 \\
&= \Big(\mathbb{E}\big[\big(\sum_a \pi_1(a \mid S_1, O_0 = 0)(Q_1^\pi(S_1, a) - \widehat{Q}_1(S_1, a))\big)^2\big]\Big)^{1/2} \\
&\le \Big(\mathbb{E}\big[\sum_a \pi_1(a \mid S_1, O_0 = 0)(Q_1^\pi(S_1, a) - \widehat{Q}_1(S_1, a))^2\big]\Big)^{1/2} \\
&:= \|Q_1^\pi - \widehat{Q}_1\|_{2,\pi},
\end{aligned}$$

where $Q_t^\pi(s,a) = \mathbb{E}(R_t + V_{t+1}^\pi \mid s,a)$. Since there is no shift on marginal distributions of states $\tilde{d}_t^\pi$ and $\tilde{d}_t^b$ when $t=1$, by Assumption 3.3,

$$\begin{aligned}
\|Q_1^\pi - \widehat{Q}_1\|_{2,\pi}^2 &= \mathbb{E}_{(S_1,O_0)\sim\tilde{d}_1^\pi}\big[\sum_a \pi_1(a \mid S_1, O_0=0)(Q_1^\pi(S_1,a) - \widehat{Q}_1(S_1,a))^2\big] \\
&= \mathbb{E}_{(S_1,O_0)\sim\tilde{d}_1^b}\big[\sum_a \pi_1(a \mid S_1, O_0=0)(Q_1^\pi(S_1,a) - \widehat{Q}_1(S_1,a))^2\big] \\
&\le \kappa_1 \mathbb{E}_{(S_1,O_0)\sim\tilde{d}_1^b}\big[\sum_a \pi_1^b(a \mid S_1)(Q_1^\pi(S_1,a) - \widehat{Q}_1(S_1,a))^2\big] \\
&:= \kappa_1 \|Q_1^\pi - \widehat{Q}_1\|_{2,\pi^b}^2,
\end{aligned}$$

and for simplicity we write $\|Q_1^\pi - \widehat{Q}_1\|_2^2$ instead of $\|Q_1^\pi - \widehat{Q}_1\|_{2,\pi^b}^2$. $\widehat{Q}_t$ is estimated from penalized nonparametric least square problem Equation (8).

$\|\widehat{Q}_t - Q_t^\pi\|_2$ involves the estimation error of the fitted Q-functions produced by FQE. Unlike standard supervised regression, the regression targets in FQE are pseudo-labels that depend on nuisance estimates and on future-stage fitted values. Concretely, at stage $t < T$, the target takes the form

$$y_{t,i} = \widehat{\widetilde{R}}_{t,i} + \widehat{V}_{t+1}^\pi(S_{t+1,i}, O_{t,i}),$$

where $\widehat{\widetilde{R}}_{t,i}$ depends on the estimated bridge $\hat{b}_t$ and $\widehat{V}_{t+1}^\pi$ depends on $\widehat{Q}_{t+1}$. Therefore, the regression noise and the regression function are statistically coupled through the common data, and a direct analysis of $\|\widehat{Q}_t - Q_t^\pi\|_2$ typically leads to non-negligible cross terms that are difficult to control without additional device such as sample splitting or cross-fitting.

To decouple the effect of nuisance estimation from the intrinsic regression error, we introduce an oracle comparator $\widehat{Q}_t^*$, defined as the solution of the same penalized regression problem as $\widehat{Q}_t$ but trained on an oracle pseudo-label y_t^*, in which the nuisance components are replaced by their population counterparts.

$$\widehat{Q}_t^* = \arg\min_{f\in\mathcal{Q}^{(t)}} \frac{1}{n}\sum_{i=1}^n \big(f(S_{t,i},A_{t,i}) - y_{t,i}^*\big)^2 + \lambda_t \|f\|_{\mathcal{Q}^{(t)}}^2, \tag{18}$$

where

$$y_{t,i}^* = \begin{cases} \widetilde{R}_{t,i}, & t = T, \\ \widetilde{R}_{t,i} + V_{t+1}^\pi(S_{t+1,i}, O_{t,i}), & t < T, \end{cases}$$

and $V_{t+1}^\pi(S_{t+1,i}, O_{t,i}) = \sum_a \pi_{t+1}(a \mid S_{t+1}, O_t) Q_{t+1}^\pi(S_{t+1}, a)$.

Then we have

$$\begin{aligned}
\|\widehat{Q}_t - Q_t^\pi\|_2 &= \|(\widehat{Q}_t - \widehat{Q}_t^*) + (\widehat{Q}_t^* - Q_t^\pi)\|_2 \\
&\le \underbrace{\|\widehat{Q}_t - \widehat{Q}_t^*\|_2}_{(a)} + \underbrace{\|\widehat{Q}_t^* - Q_t^\pi\|_2}_{(b)}.
\end{aligned}$$

For (a), since $\widehat{Q}_t$, $\widehat{Q}_t^*$ are estimated from Equation (8) and $Equation$ (18), it can be verified that

$$\|\widehat{Q}_t - \widehat{Q}_t^*\|_{2,n} \le \|y_t - y_t^*\|_{2,n} \le \sqrt{2}\|\widehat{\widetilde{R}}_t - \widetilde{R}_t\|_{2,n} + \sqrt{2}\|\widehat{V}_{t+1}^\pi - V_{t+1}^\pi\|_{2,n}.$$

Since

$$\|\widehat{\widetilde{R}}_t - \widetilde{R}_t\|_{2,n} = \|(1-O_t)(\hat{b}_t - b_t^*)\|_{2,n} \le \|\hat{b}_t - b_t^*\|_{2,n},$$

and

$$\begin{aligned}\|\widehat{V}^{\pi}_{t+1} - V^{\pi}_{t+1}\|_{2,n} :=& \Big\| \sum_a \pi_{t+1}(a \mid S_{t+1}, O_t)\big(\widehat{Q}_{t+1}(S_{t+1}, a) - Q^{\pi}_{t+1}(S_{t+1}, a)\big)\Big\|_{2,n} \\ =& \Big(\frac{1}{n}\sum_{i=1}^{n}\big[\sum_a \pi_{t+1}(a \mid S_{t+1,i}, O_{t,i})\big(\widehat{Q}_{t+1}(S_{t+1,i}, a) - Q^{\pi}_{t+1}(S_{t+1,i}, a)\big)\big]^2\Big)^{1/2} \\ \le& \Big(\frac{1}{n}\sum_{i=1}^{n}\sum_a \pi_{t+1}(a \mid S_{t+1,i}, O_{t,i})\big(\widehat{Q}_{t+1}(S_{t+1,i}, a) - Q^{\pi}_{t+1}(S_{t+1,i}, a)\big)^2\Big)^{1/2} \\ :=& \|\widehat{Q}_{t+1} - Q^{\pi}_{t+1}\|_{2,n,\pi},\end{aligned}$$

then

$$\|\widehat{Q}_t - \widehat{Q}^*_t\|_{2,n} \le \sqrt{2}\|\hat{b}_t - b^*_t\|_{2,n} + \sqrt{2}\|\widehat{Q}_{t+1} - Q^{\pi}_{t+1}\|_{2,n,\pi}. \tag{19}$$

Since $\mathcal{Q}^{(t)}$ is $(T-t+1)$-uniformly bounded, we consider scaling $\mathcal{Q}^{(t)}$ by $(T-t+1)$ for convenience.

According to Fischer & Steinwart (2020), under mild conditions, the RKHS norm of kernel ridge regression estimators is bounded with high probability. and construct RKHS ball $\mathcal{Q}^{(t)}_{R_Q} = \{Q \in \mathcal{Q}^{(t)} : \|Q\|_{\mathcal{Q}^{(t)}} \le R_Q\}$ for all $t = 1, \dots, T$. Denote difference class $\Delta\mathcal{Q}^{(t)} := \{\Delta_Q \mid \Delta_Q = Q_1 - Q_2, \quad Q_1, Q_2 \in \mathcal{Q}^{(t)}_{R_Q}\}$. Let $\bar{\delta}_{\Delta^{(t)}_Q, n}$ be the upper bound of the empirical critical radii of scaled function class $\Delta\mathcal{Q}^{(t)}$.

For RKHS $\mathcal{B}^{(t)}$, Proposition 9 in Dikkala et al. (2020) gives the closed form of the inner maximization, which implies that $\|\hat{b}_t\|_{\mathcal{B}^{(t)}}$ can be bounded by a constant R_B. Thus, consider RKHS ball $\mathcal{B}^{(t)}_{R_B} = \{b_t \in \mathcal{B}^{(t)}, \|b_t\|_{\mathcal{B}^{(t)}} \le R_B\}$. Define difference class $\Delta\mathcal{B}^{(t)} = \{\Delta_b = b_1 - b_2, \quad b_1, b_2 \in \mathcal{B}^{(t)}_{R_B}\}$ and let $\bar{\delta}_{\Delta^{(t)}_b, n}$ be the upper bound of the empirical critical radii of $\Delta\mathcal{B}^{(t)}$. Then, we define $\bar{\delta}_{\Delta_t} = \max\{\bar{\delta}_{\Delta^{(t)}_Q, n}, \bar{\delta}_{\Delta^{(t+1)}_Q, n}, \bar{\delta}_{\Delta^{(t)}_b, n}\}$ where $\delta_{\Delta_t} = \bar{\delta}_{\Delta_t} + c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$ for some $c_0, c_1 > 0$. Let $\|\widehat{Q}_{t+1} - Q^{\pi}_{t+1}\|^2_{2,b,\pi} := \mathbb{E}_{(S_{t+1}, O_t)\sim\tilde{d}^b_{t+1}}\Big[\sum_{a\in\mathcal{A}} \pi_{t+1}(a \mid S_{t+1}, O_t)\big(\widehat{Q}_{t+1}(S_{t+1}, a) - Q^{\pi}_{t+1}(S_{t+1}, a)\big)^2\Big]$.

Applying Lemma G.2 on both sides of Equation (19), we have with probability at least $1-\zeta$,

$$\begin{aligned}\|\widehat{Q}_t - \widehat{Q}^*_t\|_2 \lesssim& \|\hat{b}_t - b^*_t\|_2 + \|\widehat{Q}_{t+1} - Q^{\pi}_{t+1}\|_{2,b,\pi} + (T-t+1)\delta_{\Delta_t} \\ (\textit{Assumption 3.3 (2)}) \lesssim& \|\hat{b}_t - b^*_t\|_2 + \sqrt{\kappa_{t+1}}\|\widehat{Q}_{t+1} - Q^{\pi}_{t+1}\|_2 + (T-t+1)\delta_{\Delta_t}.\end{aligned} \tag{20}$$

(b) corresponds to a standard penalized least square estimation error. Since $\|Q^*_t\|_{\mathcal{Q}^{(t)}}$ is bounded, by Lemma G.7, with probability at least $1-\zeta$, (b) is bounded by

$$\|\widehat{Q}^*_t - Q^{\pi}_t\|_2 \lesssim (\delta_{\Delta^{(t)}_Q} + \sqrt{\lambda_t})(T-t+1),$$

where $\delta_{\Delta^{(t)}_Q} = \delta_{\Delta^{(t)}_Q, n} + c_0\sqrt{\frac{\log(c_1 T/\zeta)}{n}}$ for some $c_0, c_1 > 0$.

Therefore, with probability at least $1 - \zeta/T$,

$$\begin{aligned}\|\widehat{Q}_t - Q^{\pi}_t\|_2 \le& \|\widehat{Q}_t - \widehat{Q}^*_t\|_2 + \|\widehat{Q}^*_t - Q^{\pi}_t\|_2 \\ \lesssim& \|\hat{b}_t - b^*_t\|_2 + \sqrt{\kappa_{t+1}}\|\widehat{Q}_{t+1} - Q^{\pi}_{t+1}\|_2 + (T-t+1)\delta_{\Delta_t} + (\delta_{\Delta^{(t)}_Q} + \sqrt{\lambda_t})(T-t+1).\end{aligned}$$

Applying backward induction from $t = T$ down to $t = 1$ yields the bound for (II) with probability at least $1 - \zeta$:

$$\begin{aligned}
|\mathbb{E}[V_1^\pi] - \mathbb{E}[\widehat{V}_1^\pi]| \leq & \|V_1^\pi - \widehat{V}_1^\pi\|_2 \\
\leq & \sqrt{\kappa_1}\|\widehat{Q}_1 - Q_1^\pi\|_2 \\
\lesssim & \sum_{t=1}^{T} \Big(\prod_{j=1}^{T} \sqrt{\kappa_j} \Big) \Big[\|\hat{b}_t - b_t^*\|_2 + (T-t+1)\delta_{\Delta_t} + (\delta_{\Delta_Q^{(t)}} + \sqrt{\lambda_t})(T-t+1) \Big] \\
\lesssim & K \sum_{t=1}^{T} \Big[\tau_t \delta_t (1 + \|b_t^*\|^2_{\mathcal{B}^{(t)}}) + (\delta_{\Delta_t} + \delta_{\Delta_Q^{(t)}} + \sqrt{\lambda_t})(T-t+1) \Big].
\end{aligned}$$

F.4. Bound of (III)

For (III), we first define function class $\Delta\mathcal{V}^{(t)} = \big\{\Delta = V_1 - V_2 \mid V_1, V_2 \in \mathcal{V}_{R_V}^{(t)}\big\}$, where $\mathcal{V}_{R_V}^{(t)}$ is a $(T-t+1)$-uniformly bounded function class of value functions at time t, induced from $\mathcal{Q}^{(t)}$ under operator Π_t: $\mathcal{V}_{R_V}^{(t)} = \{\Pi_t Q : Q \in \mathcal{Q}_{R_Q}^{(t)}\}$. Here linear operator Π_t is defined as $(\Pi_t Q)(s, o-) = \sum \pi_t(a \mid s, o_-)Q(s, a)$. We choose the cost function as $\mathcal{L}(f(X), Y) = f(X)$ and apply Lemma G.1. We here also scale $\mathcal{V}_{R_V}^{(t)}$ by $(T-t+1)$.

Then, with probability at least $1 - \zeta$,

$$|\mathbb{E}(V_1^\pi - \widehat{V}_1^\pi) - \mathbb{E}_n(V_1^\pi - \widehat{V}_1^\pi)| \lesssim \delta_{\Delta_V^{(1)}}(\|V_1^\pi - \widehat{V}_1^\pi\|_2 + T\delta_{\Delta_V^{(1)}}),$$

where $\delta_{\Delta_V^{(1)}} = \bar{\delta}_{\Delta_V^{(1)},n} + c_0\sqrt{\frac{\log(c_1/\zeta)}{n}}$, and $\bar{\delta}_{\Delta_V^{(1)},n}$ is the upper bound of the empirical critical radii of scaled function class $\Delta\mathcal{V}^{(1)}$. Note that $\delta_{\Delta_V^{(1)}} \leq \delta_{\Delta_Q^{(1)}}$.

F.5. Policy value error bound

Combine the above inequalities, we obtain the policy value estimation error bound with probability at least $1 - \zeta$:

$$\begin{aligned}
\big|\widehat{V}(\pi) - V(\pi)\big| \leq & (I) + (II) + (III) \\
\lesssim & T\sqrt{\frac{\log(c_1 T/\zeta)}{n}} + (\delta_{\Delta_V^{(1)}} + 1)\|V_1^\pi - \widehat{V}_1^\pi\|_2 + T(\delta_{\Delta_V^{(1)}})^2 \\
\lesssim & T\sqrt{\frac{\log(c_1 T/\zeta)}{n}} + T(\delta_{\Delta_V^{(1)}})^2 + K(\delta_{\Delta_V^{(1)}} + 1)\Big[\tau_t\delta_t(1 + \|b_t^*\|^2_{\mathcal{B}^{(t)}}) + \\
& (\delta_{\Delta_t} + \delta_{\Delta_Q^{(t)}} + \sqrt{\lambda_t})(T-t+1)\Big] \\
\lesssim & T\sqrt{\frac{\log(c_1 T/\zeta)}{n}} + T(\delta_{\Delta_V^{(1)}})^2 + K(\delta_{\Delta_V^{(1)}} + 1)\Big[\tau_t\delta_t(1 + \|b_t^*\|^2_{\mathcal{B}^{(t)}}) + \\
& (T-t+1)\delta_{t,*}\Big] \\
\lesssim & T\sqrt{\frac{\log(c_1 T/\zeta)}{n}} + K\tau_{\max} T \sum_{t=1}^{T} \delta_{t,*},
\end{aligned}$$

where $\delta_{t,*}$ as the maximum of the critical radii of difference classes $\Delta\mathcal{Q}^{(t)}$, $\Delta\mathcal{Q}^{(t+1)}$, $\Delta\mathcal{B}^{(t)}$, and $\mathcal{G}_U^{(t)}$ for $t = 1, \ldots, T$, namely $\delta_{\Delta_Q^{(t)}}$, $\delta_{\Delta_Q^{(t+1)}}$, $\delta_{\Delta_B^{(t)}}$ and $\delta_{G^{(t)}}$.

With polynomial decay

$$\mu_{t,j}^Q \lesssim j^{-2\alpha_Q}, \quad \mu_{t,j}^B \lesssim j^{-2\alpha_B}, \quad \mu_{t,j}^G \lesssim j^{-2\alpha_G}, \quad \alpha_Q, \alpha_B, \alpha_G > 1/2,$$

the corresponding critical radii satisfy

$$\delta_{\Delta_Q^{(t)}}, \delta_{\Delta_Q^{(t+1)}} \lesssim R_Q^{\frac{1}{2\alpha_Q+1}} n^{-\frac{\alpha_Q}{2\alpha_Q+1}} \log n,$$
$$\delta_{\Delta_B^{(t)}} \lesssim R_B^{\frac{1}{2\alpha_B+1}} n^{-\frac{\alpha_B}{2\alpha_B+1}} \log n,$$
$$\delta_{G^{(t)}} \lesssim U_t^{\frac{1}{2\alpha_G+1}} n^{-\frac{\alpha_G}{2\alpha_G+1}} \log n.$$

Thus, the critical radius $\delta_{t,*}$ satisfies

$$\delta_{t,*} \lesssim \max\{\sqrt{R_Q}, \sqrt{R_B}, \sqrt{U_t}\} n^{-\frac{\alpha_{\min}}{2\alpha_{\min}+1}} \log n,$$

where $\alpha_{\min} = \min\{\alpha_Q, \alpha_B, \alpha_G\}$. Therefore, with probability at least $1-\zeta$, the policy value is bound by

$$\left|\widehat{V}(\pi) - V(\pi)\right| \lesssim K\tau_{\max} T^2 \sqrt{\log(c_1 T/\zeta)} n^{-\frac{\alpha_{\min}}{2\alpha_{\min}+1}} \log n.$$

G. Auxiliary lemmas

Lemma G.1 (Wainwright (2019), Theorem 14.20)**.** *Suppose function class $\mathcal{F}$ is symmetric, 1-uniformly bounded, and star-shaped around f^*. Let $\delta_n^2 \geq \frac{c}{n}$ be any solution to the inequality $\mathcal{R}_n(\mathcal{F}^*, \delta) \leq \delta^2$, where $\mathcal{F}^* = \{f - f^* \mid f \in \mathcal{F}\}$. Suppose the cost function $\mathcal{L}(f(X), Y)$ is L-Lipschitz in its first argument $f(X)$. Then for all $f \in \mathcal{F}$, with probability at least $1 - c_1 e^{-c_2 n \delta_n^2}$, we have*

$$|\mathbb{E}_n\big(\mathcal{L}(f(x), y) - \mathcal{L}(f^*(x), y)\big) - \mathbb{E}\big(\mathcal{L}(f(x), y) - \mathcal{L}(f^*(x), y)\big)| \leq 10L\delta_n\big(\|f - f^*\|_2 + \delta_n\big).$$

Lemma G.2 (Wainwright (2019), Theorem 14.1)**.** *Given a star-shaped and b-uniformly bounded function class $\mathcal{F}$, set $\delta_n > 0$ be any solution to $\mathcal{R}(\mathcal{F}, \delta) \leq \frac{\delta^2}{b}$. Then for any $t \geq \delta_n$, with probability at least $1 - c_1 \exp(-c_2 \frac{nt^2}{b^2})$, we have*

$$\big|\|f\|_{2,n}^2 - \|f\|_2^2\big| \leq \frac{1}{2}\|f\|_2^2 + \frac{1}{2}t^2$$

for all $f \in \mathcal{F}$.

Lemma G.3 (Foster & Syrgkanis (2023), Lemma 14)**.** *Consider a 1-uniformly bounded and star-shaped function class $\mathcal{F}$, and pick any $f^* \in \mathcal{F}$. Let $\delta_n^2 \geq c_1 \frac{\log(\log n)}{n}$ be any solution to the inequalities $\mathcal{R}_n(\mathcal{F}_t^*, \delta) \leq \delta^2$ for all $t \in \{1, \ldots, d\}$, where $\mathcal{F}_t^* = \{f_t - f_t^* \mid f_t \in \mathcal{F}|_t\}$. Assume $\mathcal{L}_f$ is L-Lipschitz in its first argument f with respect to its ℓ_2 norm. Then for all $f \in \mathcal{F}$, for some universal constants $c_2, c_3 > 0$, with probability at least $1 - c_2 e^{-c_3 n \delta_n^2}$, we have*

$$|\mathbb{E}_n\big(\mathcal{L}_f - \mathcal{L}_{f^*}\big) - \mathbb{E}\big(\mathcal{L}_f - \mathcal{L}_{f^*}\big)| \leq 18Ld\delta_n\big(\|f - f^*\|_2 + \delta_n\big).$$

The outcome $\hat{f}$ of constrained ERM satisfies that with the same probability,

$$\mathbb{E}_n\big(\mathcal{L}_{\hat{f}} - \mathcal{L}_{f^*}\big) \leq 18Ld\delta_n\big(\|\hat{f} - f^*\|_2 + \delta_n\big).$$

Lemma G.4 (Wainwright (2019), Example 3.5)**.** *Let $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ be i.i.d. Rademacher variables taking values in $\{-1, +1\}$ with equal probability. Let $\mathcal{A} \subset \mathbb{R}^n$ be any (possibly infinite) bounded set, and define*

$$Z(\mathcal{A}) := \sup_{a \in \mathcal{A}} \langle a, \varepsilon \rangle = \sup_{a \in \mathcal{A}} \sum_{k=1}^{n} a_k \varepsilon_k.$$

Let $W(\mathcal{A}) := \sup_{a \in \mathcal{A}} \|a\|_2$. Then for all $t > 0$,

$$\mathbb{P}\Big(Z(\mathcal{A}) \geq \mathbb{E}[Z(\mathcal{A})] + t\Big) \leq \exp\Big(-\frac{t^2}{16\, W(\mathcal{A})^2}\Big).$$

Moreover, since $-Z(\mathcal{A}) = \inf_{a \in \mathcal{A}} \langle a, \varepsilon \rangle$ and the same argument applies,

$$\mathbb{P}\Big(|Z(\mathcal{A}) - \mathbb{E}[Z(\mathcal{A})]| \geq t\Big) \leq 2\exp\Big(-\frac{t^2}{16\, W(\mathcal{A})^2}\Big).$$

Lemma G.5 (Wainwright (2019), Corollary 14.5). *Let $\mathcal{H}$ be an RKHS with reproducing kernel K and let $\mathcal{F} := \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$ be the unit ball. Let $\{\mu_j\}_{j=1}^{\infty}$ denote the non-increasing eigenvalues. Then the local Rademacher complexity satisfies, for any $\delta > 0$,*

$$\mathcal{R}_n(\mathcal{F}, \delta) \leq \sqrt{\frac{2}{n}} \left(\sum_{j=1}^{\infty} \min\{\mu_j, \delta^2\} \right)^{1/2}.$$

Moreover, let $\{\hat{\mu}_j\}_{j=1}^{n}$ denote the eigenvalues of the renormalized kernel matrix $\mathbf{K} \in \mathbb{R}^{n \times n}$ with entries $\mathbf{K}_{ij} = K(x_i, x_j)/n$. Then the local empirical Rademacher complexity satisfies, for any $\delta > 0$,

$$\widehat{\mathcal{R}}_n(\mathcal{F}, \delta) \leq \sqrt{\frac{2}{n}} \left(\sum_{j=1}^{n} \min\{\widehat{\mu}_j, \delta^2\} \right)^{1/2}.$$

Lemma G.6 (Krieg (2018), Theorem 1(i)). *Let $\sigma : \mathbb{N} \to \mathbb{R}_+$ be a non-increasing sequence with $\sigma(n) \to 0$. For $d \in \mathbb{N}$, define its d-th tensor power*

$$\sigma_d(n_1, \ldots, n_d) = \prod_{k=1}^{d} \sigma(n_k), \quad (n_1, \ldots, n_d) \in \mathbb{N}^d,$$

and let $\tau : \mathbb{N} \to \mathbb{R}_+$ be the non-increasing rearrangement of $\{\sigma_d(n_1, \ldots, n_d)\}_{(n_1,\ldots,n_d) \in \mathbb{N}^d}$. If for some $s > 0$ one has $\sigma(n) \lesssim n^{-s}$, then

$$\tau(n) \lesssim n^{-s} (\log n)^{s(d-1)}.$$

Lemma G.7 (Rademacher analogue of Wainwright (2019), Theorem 13.17). *Let $(x_i, y_i)_{i=1}^{n}$ be i.i.d. with $y_i = f^*(x_i) + \xi_i$, where $\mathbb{E}[\xi_i \mid x_i] = 0$ and ξ_i is conditionally σ-sub-Gaussian. Let $\mathcal{F}$ be a symmetric, star-shaped class equipped with a Hilbert norm $\|\cdot\|_{\mathcal{F}}$. Consider the penalized least squares estimator*

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathcal{F}}^2 \right\}.$$

Suppose $f^* \in \mathcal{F}$ and $\|f^*\|_{\mathcal{F}} \leq R$. Define the localized difference class

$$\mathcal{G} := \{\Delta = f - f^* : f \in \mathcal{F},\ \|f\|_{\mathcal{F}} \leq R\}.$$

and the local empirical Rademacher complexity

$$\widehat{\mathcal{R}}_n(\mathcal{G}, \delta) := \mathbb{E}_{\varepsilon} \left[\sup_{\Delta \in \mathcal{G} : \|\Delta\|_{2,n} \leq \delta} \left| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \Delta(x_i) \right| \,\middle|\, x_{1:n} \right].$$

Let $\bar{\delta}_n$ be the upper bound of the critical radii satisfying $\widehat{\mathcal{R}}_n(\mathcal{G}, \bar{\delta}_n) \leq \frac{\bar{\delta}_n^2}{32\sigma}$, and define $\delta_n = \bar{\delta}_n + c_0 \sigma \sqrt{\frac{\log(c_1/\zeta)}{n}}$ for some numerical constants $c_0, c_1 > 0$. Assume that $\lambda_n \geq \frac{3}{4}\delta_n^2$, then there exist constant $C_1 > 0$ such that, with probability at least $1 - \zeta$,

$$\|\hat{f} - f^*\|_2^2 \leq C_1 R^2 \big(\delta_n^2 + \lambda_n\big).$$