Model-based Reinforcement Learning for Confounded POMDPs

Mao Hong [1] **Zhengling Qi** [2] **Yanxun Xu** [1]

Abstract

We propose a model-based offline reinforcement learning (RL) algorithm for confounded partially observable Markov decision processes (POMDPs) under general function approximations and show it is provably efficient under some technical conditions such as the partial coverage imposed on the offline data distribution. Specifically, we first establish a novel model-based identification result for learning the effect of any action on the reward and future transitions in the confounded POMDP. Using this identification result, we then design a nonparametric two-stage estimation procedure to construct an estimator for off-policy evaluation (OPE), which permits general function approximations. Finally, we learn the optimal policy by performing a conservative policy optimization within the confidence regions based on the proposed estimation procedure for OPE. Under some mild conditions, we establish a finite-sample upper bound on the suboptimality of the learned policy in finding the optimal one, which depends on the sample size and the length of horizons polynomially.

1. Introduction

Reinforcement Learning (RL) (Sutton & Barto, 2018) has been recognized as an effective approach for optimizing sequential decision-making processes. It seeks to learn an optimal policy by maximizing the expected cumulative rewards. However, most existing literature has focused on environments that are fully observable with Markovian transition dynamics, which may not be known a priori. In practice, the challenge of partial observability of state information frequently arises, making the Markov decision processes (MDPs) unsuitable for modeling the underlying data-generating processes. For example, in autonomous driving, the environment is typically not fully observed. Instead, only partial information, such as noisy images or videos captured by cameras, is available (Sun et al., 2020). Considering partial observability inherent in many applications, partially observable Markov decision processes (POMDPs) (Monahan, 1982) are considered as a more appropriate framework for sequential decision-making for a wide range of applications (e.g., Sawaki & Ichikawa, 1978; Albright, 1979; Monahan, 1982; Singh et al., 1994; Jaakkola et al., 1994; Cassandra, 1998; Young et al., 2013; Zhang & Bareinboim, 2016; Bravo et al., 2019). Moreover, in light of the ethical and logistical challenges faced by online learning such as the assignment of patients to potentially inferior or harmful treatments in healthcare (Gottesman et al., 2019), offline RL emerged and has recently received a lot of research interests (Levine et al., 2020). In the offline setting, an agent aims to perform policy evaluation and learning by only using a pre-collected dataset, which may be more practical in solving decision-making problems in some high stake domains.

Due to these practical challenges, there is a recent line of research focusing on developing offline RL methods for confounded POMDPs (e.g., Tennenholtz et al., 2020; Bennett & Kallus, 2021; Nair & Jiang, 2021; Shi et al., 2022; Miao et al., 2022; Lu et al., 2022; Hong et al., 2023). The confounding effect (Pearl, 2009) in this context arises from the offline data-generating processes, wherein the behavior policy depends on the unobserved states. In this setting, unobserved state variables act as unmeasured confounders at each decision point, which can simultaneously affect the action, the reward, and the future transition. This complexity introduces a confounding bias when standard offline RL methods designed for MDPs fail. To address this issue, some aforementioned works employ proxy variables for policy evaluation and learning. A significant portion of these investigations focuses on the task of off-policy evaluation (e.g., Tennenholtz et al., 2020; Bennett & Kallus, 2021; Nair & Jiang, 2021; Shi et al., 2022; Miao et al., 2022), with only a few exploring the problem of offline policy learning (Lu et al., 2022; Hong et al., 2023). In particular, Hong et al. (2023) introduced a first policy gradient method for confounded POMDPs, but under some stringent *full coverage*

[1]Department of Applied Mathematics and Statistics, Johns Hopkins University, Baltimore, United States [2]Department of Decision Sciences, George Washington University, Washington, DC, United States. Correspondence to: Zhengling Qi <qizhengling@gwu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

assumption, which requires that the offline data can cover trajectories generated by all policies - a condition that is often hard to verify. On the other hand, to relax the full coverage assumption, Lu et al. (2022) adopted a pessimistic principle (Xie et al., 2021; Uehara & Sun, 2021; Fu et al., 2022, e.g.,), which is commonly applied in offline RL methods for standard MDPs, and extend it to POMDPs. Thus, their algorithm can find an in-class optimal policy by only requiring *partial coverage*, meaning that the offline data only needs to cover the trajectories of the optimal policy. Nonetheless, Lu et al. (2022)'s exploration of this problem is restricted to a *model-free* context, leaving a gap in *model-based* RL algorithms for confounded POMDPs under the partial coverage assumption. Motivated by these, this paper aims to bridge this gap by proposing a *model-based* RL method for confounded POMDPs.

In this article, we introduce a novel offline model-based policy learning method for confounded POMDPs with continuous state and observation spaces under partial coverage. Specifically, we first establish a novel nonparametric identification for the policy value of any history-dependent policy from a model-based viewpoint. This identification enables consistent estimation of policy values using only observable offline data, eliminating the confounding bias introduced by partial observability. Furthermore, our approach accommodates general function approximation, which is necessary given continuous state spaces and observation spaces. Following the identification, which entails solving a sequence of conditional moment equations, we design a nonparametric two-stage estimation procedure to estimate the policy value. Finally, to relax the full coverage assumption, we extend the principle of pessimism to confounded POMDPs tailored for our model-based structure. In particular, we construct a series of confidence regions based on the estimation procedure and perform a conservative policy optimization within the constructed confidence regions, i.e. learning a policy that aims to maximize the most pessimistic estimator of the policy value within confidence regions.

We summarize the main technical challenges addressed in this paper. (1) Developing model-based policy value identification under confounded POMDPs presents a significant challenge due to the unobservable states/confounders, which make it hard to extract the information of reward and transition model from the offline dataset. (2) In the estimation process of the bridge function b, we encounter a new type of conditional moment restriction problem, which deviates from existing methods. This new challenge motivates the development of new estimating approaches and theoretical analysis, distinguishing our work from existing literature. (3) In establishing an upper bound on the suboptimality, we face the challenge of decomposing the suboptimality into one-step errors which comes from the estimation of reward and transition models.

Our contributions can be summarized as follows. First, to the best of our knowledge, the proposed policy value identification for confounded POMDPs is the first result under the function approximation settings within the model-based framework. While Tennenholtz et al. (2020) proposed a model-based identification in the tabular setting, their methods are not applicable to settings with continuous observation/state spaces. Compared to the existing model-free methods such as Shi et al. (2022); Bennett & Kallus (2021); Miao et al. (2022), our theoretical derivation is novel. In particular, they typically rely on solving a series of Bellman-type backward moment equations to directly identify the policy value, while our approach emphasizes the extraction of information from both reward and transition models, which are independent of the policy. Due to this intrinsic characteristic of our model-based framework, as a by-product, the marginal distribution of the cumulative reward induced by any policy can also be identified and more efficiently computed, compared with model-free methods. See Remark 3.6 for more details.

Secondly, we introduce a nonparametric two-stage estimation method aimed at solving a new-type conditional moment restriction problem, i.e., estimating the function b by solving $\mathbb{E}_W[b(W, y) \mid X] = p(y \mid X)$ for every y, where W, X are generic random vectors and $p(y \mid X)$ denotes the conditional density function. Note that this is different from the standard conditional moment restriction problem as y is deterministic and one needs to solve it simultaneously across all y for obtaining the target function b. The exploration of this problem in the existing literature is scarce and there is little theoretical result. In this paper, we formalize a valid risk functional, based on which we design the estimation method, and establish the corresponding theoretical guarantee.

Finally, we demonstrate the validity of the proposed algorithm by providing a finite-sample upper bound of the performance between the learned policy and the optimal policy under the partial coverage assumption. In particular, a novel theoretical derivation for decomposing the differences of the true policy value and the estimated policy value into a polynomial function of key parameters and error terms (Lemma C.1) could be of independent interest. Moreover, we provide a sharp contrast between the proposed model-based and those model-free approaches such as (Lu et al., 2022) for confounded POMDPs. Notably, our proposed model-based method does not have restriction on the policy space, compared with the model-free methods in this setting. This is particularly appealing when the global optimal policy is not contained in the pre-specified policy class. See Remarks 3.7 and 4.3 for more details on the comparisons.

2. Preliminaries

Consider an episodic and finite-horizon POMDP denoted by $\mathcal{M} := (\mathcal{S}, \mathcal{O}, \mathcal{A}, T, \nu_1, \{P_t\}_{t=1}^T, \{\mathcal{E}_t\}_{t=1}^T, \{r_t\}_{t=1}^T)$, where $\mathcal{S}$, $\mathcal{O}$ and $\mathcal{A}$ denote the state, observation and action spaces respectively. Without loss of generality, we assume that both $\mathcal{S}$ and $\mathcal{O}$ are *continuous*, while the action space $\mathcal{A}$ is *finite*. The integer T is set as the total length of the horizon. We use $\nu_1 \in \Delta(\mathcal{S})$ to denote the distribution of the initial state, where $\Delta(\mathcal{S})$ is a class of all probability distributions over $\mathcal{S}$. In addition, we denote $\{P_t\}_{t=1}^T$ by the collection of state transition kernels over $\mathcal{S} \times \mathcal{A}$ to $\mathcal{S}$, and $\{\mathcal{E}_t\}_{t=1}^T$ by the collection of observation emission kernels over $\mathcal{S}$ to $\mathcal{O}$. Lastly, we use $\{r_t\}_{t=1}^T$ to denote the collection of reward functions, i.e., $r_t : \mathcal{S} \times \mathcal{A} \to [-1, 1]$ at each decision point t. In a standard POMDP, at each decision point t, $O_t \sim \mathcal{E}_t(\cdot \mid S_t)$ is observed given the current (hidden) state S_t. Then the agent selects an action A_t following some policy, and receives an immediate reward R_t with $\mathbb{E}[R_t \mid S_t = s, A_t = a] = r_t(s, a)$ for every (s, a). The system then transits to the next state S_{t+1} according to some transition kernel $P_t(\cdot \mid S_t, A_t)$. Thus the underlying dynamics follows MDP. The corresponding directed acyclic graph (DAG) is depicted in Figure 1. Different from an MDP, the state variable S_t cannot be observed in a POMDP.

The goal of this paper is to find an optimal history-dependent policy for POMDPs. Define the observed history up to the decision point t by $H_t := (O_1, A_1, ..., O_t, A_t) \in \mathcal{H}_t$, where $\mathcal{H}_t := \prod_{j=1}^t \mathcal{O} \times \mathcal{A}$ is the corresponding space. Then at each t, the history-dependent policy π_t is defined as a function mapping from $\mathcal{O} \times \mathcal{H}_{t-1}$ to $\Delta(\mathcal{A})$. For any generic policy $\pi = \{\pi_t\}_{t=1}^T$, the corresponding value is defined as

$$\mathcal{V}(\pi) := \mathbb{E}^\pi[\sum_{t=1}^T R_t \mid S_1 \sim \nu_1],$$

where $\mathbb{E}^\pi$ is taken with respect to the distribution such that all actions are determined by the policy π. In this work, we aim to develop a model-based RL algorithm to find an optimal policy π^* defined as,

$$\pi^* \in \arg\max_\pi \mathcal{V}(\pi).$$

In the offline setting, we assume that a decision maker can only access a pre-collected dataset, which is generated by some behavior policy $\{\pi_t^b\}_{t=1}^T$, but unable to further interact with the environment. The behavior policy considered in this work could possibly depend on the unobserved state S_t, i.e., $\pi_t^b : \mathcal{S} \to \Delta(\mathcal{A})$ for each t, which makes our problem more challenging compared with online POMDPs. We use $\mathbb{P}^{\pi^b}$ to denote the offline data distribution and summarize the data as $\mathcal{D} := (o_t^n, a_t^n, r_t^n)_{t=1:T}^{n=1:N}$, which are N i.i.d. copies from $\mathbb{P}^{\pi^b}$.

To develop a model-based algorithm for finding an optimal policy π^* using the offline data, one needs to identify and compute the effect of actions on the immediate reward and future transitions. Once the dynamic is learned, a pessimistic model-based RL algorithm can be implemented to learn the optimal policy. To proceed with this idea, there are two main challenges: (1) estimating the reward and future transitions based on the action with function approximations only using the offline dataset $\mathcal{D}$ and (2) developing an algorithm with theoretical guarantee for finding an optimal policy under the partial coverage assumption. The first challenge lies in that the state variable S_t is unobserved and the history-dependent transition dynamics conditioning on all past actions may not be identified by the offline data. Furthermore, function approximations are needed when both state and observation spaces are continuous. The second challenge involves developing a valid confidence set to quantify the uncertainty associated with estimating the dynamics using the offline data.

Notations. Throughout this paper, we assume that $\mathbb{E}$ is taken with respect to the offline distribution. Similarly, we use the notation $X \perp\!\!\!\perp Y \mid Z$ when X and Y are conditionally independent given Z under the offline distribution. For any two sequences $\{a_n\}_{n=1}^\infty$, $\{b_n\}_{n=1}^\infty$, $a_n \lesssim b_n$ denotes $a_n \leq C b_n$ for some $N, C > 0$ and every $n > N$. If $a_n \lesssim b_n$ and $b_n \lesssim a_n$, then $a_n \asymp b_n$. Big O and O_P are used as conventions. For any policy π that depends on the observed data, the suboptimality gap is defined as

$$\mathrm{SubOpt}(\pi) := \mathcal{V}(\pi^\star) - \mathcal{V}(\pi).$$

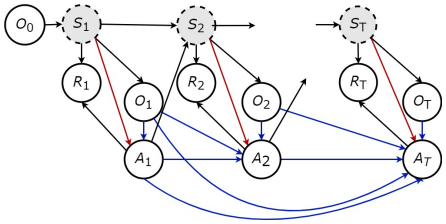

Figure 1. The directed acyclic graph illustrates the data generating process in confounded POMDPs, where states S_t are unobserved. Red arrows represent the generation of actions via the behavior policy, while blue arrows denote the generation through a history-dependent policy.

3. Methods

In this section, we introduce the proposed model-based RL method for confounded POMDPs.

In Section 3.1, we first establish a novel model-based policy value identification result for addressing the issue of confounding bias caused by partial observability in POMDPs

with continuous state and observation spaces. In particular, the policy value of any policy can be identified via a series of reward-emission bridge functions and a series of dynamic-emission bridge functions which are solutions to a sequence of conditional moment restrictions. Then, inspired by Singh et al. (2019); Mastouri et al. (2021), we develop a two-stage nonparametric estimation procedure in Section 3.2 for estimating the required bridge functions based on these conditional moment restrictions. Such nonparametric estimation procedure will allow general function approximations for estimating bridge functions and subsequently for estimating policy values. We remark that function approximation is inevitable when state and observation spaces are continuous. Finally, in Section 3.3, we incorporate the pessimism principle into our model-based method to handle the issue of distribution shift in the offline setting. In particular, we perform a conservative policy optimization within two confidence regions of bridge functions so that the learned policy is restricted within the offline data distribution and does not induce over-exploration.

3.1. Policy Value Identification

Since the observed decision process does not satisfy the Markov property, standard off-policy evaluation methods developed for MDPs cannot be applied. This becomes more challenging when the behavior policy could depend on the hidden states as well. In this case, at each decision point t, S_t will confound the effect of action A_t on the immediate reward and all future transitions. Without taking this into account will lead to bias estimation of the policy value (Hong et al., 2023).

To address this confounding bias, we establish a novel policy value identification result from the *model-based* perspective for confounded POMDPs when both state and observation spaces are continuous. To start with, we impose several standard assumptions on the data-generating process under the framework of POMDPs. See Figure 1. In addition, we restrict our study to a class of POMDPs in which the information of unobserved states can be captured by the observed variables in the dataset.

Same as many existing works on POMDP (Shi et al., 2022; Miao et al., 2022; Lu et al., 2022), we first assume the availability of some baseline covariates, represented by O_0, which carry some information before the decision-making process. The initial data for all individuals can be recorded as $\{o_0^n\}_{n=1}^N$. To enable the observable trajectory $\{O_t\}_{t=0}^T$ for identifying the policy value, we impose Assumption 3.1 in the following.

Assumption 3.1. It holds that

$$O_0 \perp\!\!\!\perp (O_t, O_{t+1}, R_t) \mid S_t, A_t, H_{t-1},\ \forall t = 1, ..., T. \quad (1)$$

Assumption 3.1 is a mild condition when O_0 is pre-collected before the decision process. To identify the policy value, we assume the existence of certain *bridge functions*, which are summarized in the following assumption.

Assumption 3.2 (Existence of bridge functions)**.** There exist reward-emission bridge functions $\{b_R^{[t]} : \mathcal{A} \times \mathcal{O} \times [-1, 1] \times \mathcal{O} \to \mathbb{R}\}_{t=1}^T$ and dynamic-emission bridge functions $\{b_D^{[t]} : \mathcal{A} \times \mathcal{O} \times \mathcal{O} \times \mathcal{O} \to \mathbb{R}\}_{t=1}^{T-1}$ that satisfy the following conditional moment restrictions for each $t = 1, ..., T$:

$$\begin{aligned} &\mathbb{E}[b_R^{[t]}(A_t, O_t, r_t, o_t) \mid H_{t-1}, A_t, O_0] \\ =&p(r_t, o_t \mid H_{t-1}, A_t, O_0), \text{ and,} \end{aligned} \quad (2)$$

$$\begin{aligned} &\mathbb{E}[b_D^{[t]}(A_t, O_t, o_{t+1}, o_t) | H_{t-1}, A_t, O_0] \\ =&p(o_{t+1}, o_t \mid H_{t-1}, A_t, O_0). \end{aligned} \quad (3)$$

The existence of such bridge functions is justified by some mild regularity conditions of conditional expectation operators $\mathbb{E}_{A_t, O_t | A_t, H_{t-1}, O_0}$ with tools from singular value decomposition in functional analysis (Kress et al., 1989, Chapter 15). See also Appendix B.1 of Hong et al. (2023) for an instantiation of required regularity conditions in confounded POMDPs.

Similar versions of Assumption 3.2 have also been utilized in one recently developed causal inference method called double negative control (Miao et al., 2018; Tchetgen et al., 2020) and off-policy evaluation methods in confounded POMDPs with continuous state and observation spaces in the *model-free* settings (Bennett & Kallus, 2021; Shi et al., 2022; Miao et al., 2022). In this work, we study the *model-based* counterpart along with the following completeness assumption for identifying the policy value.

Assumption 3.3 (Completeness)**.** For any measurable function $g_t : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and any $1 \le t \le T$,

$$\mathbb{E}[g_t(S_t, A_t) \mid A_t, H_{t-1}, O_0] = 0$$

almost surely if and only if $g_t(S_t, A_t) = 0$ almost surely.

Assumption 3.3 essentially requires that the observed $\{O_0, H_{t-1}\}$ carries sufficient information of the unobserved S_t. There are many commonly-used statistical and econometric models which satisfy Assumption 3.3. Examples include exponential families (Newey & Powell, 2003) and location-scale families (Hu & Shiu, 2018). The completeness assumption is also widely made to ensure the uniqueness of instrumental variable estimation. See Newey & Powell (2003) for more details.

Remark 3.4. The main difference between our assumptions and those in the model-free settings (Shi et al., 2022; Lu et al., 2022; Hong et al., 2023) lies in Assumption 3.2. In this paper, bridge functions depend on (r_t, o_t) and (o_{t+1}, o_t) (compared to the model-free counterpart), and the equations need to hold for all (r_t, o_t) and all (o_{t+1}, o_t). In contrast,

the counterpart assumption in model-free settings additionally depends on the policy π_θ (compared to the model-based version), and the equations need to hold for all π_θ. Intuitively, when the policy space is very large, the model-based framework, as delineated by Assumption 3.2, may offer a more feasible approach due to its simpler validation requirements and a theoretical upper bound independent of the policy space size. Conversely, when the policy space is rather small, and the reward/state space is very large, the model-free framework might present advantages due to the lower complexity in satisfying the bridge function's existence assumption.

Assumptions 3.2 and 3.3 imply that the confounding effect of each action due to the unobservable state S_t on the bridge function matches that on the outcome of interest, i.e., the current reward, the current weight, and the next state. Hence the bridge function can be used as a good "substitute." Moreover, thanks to the conditional independence of A_t given S_t, the bridge function can correct the bias and eventually identify the policy value. It is known that solving a general POMDP is NP-hard (Burago et al., 1996; Vlassis et al., 2012). In this paper, we navigate away from the NP-hard complexity by focusing on a more manageable problem class—a learnable subclass of POMDPs that satisfies Assumptions 3.2 and 3.3.

A medical application scenario can be introduced to further illustrate the assumptions listed above. The latent state S_t represents some clinical state of a patient, while the observable variable O_t corresponds to data accessible to a physician through medical diagnostics, reflecting the patient's state S_t. In this context, A_t denotes the administered treatment. According to Assumptions 3.1, 3.2, 3.3, our strategy involves selecting observations $\{O_t\}_{t=1}^T$ (e.g. blood pressure, heart rate) such that both O_t and the history H_{t-1} contain sufficient information to reflect the latent state S_t.

Finally, the key identification results for policy value under the model-based perspective are summarized in Theorem 3.5. A detailed proof is provided in Appendix B.

Theorem 3.5 (Main identification results)**.** *Under Assumptions 3.1, 3.2, 3.3, for each $t = 1, ..., T$, $p^\pi(r_t)$ can be identified by*

$$p^\pi(r_t) = \int_{h_t} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, h_{j-1}) f_t(r_t, h_t), \tag{4}$$

where $f_t(r_t, h_t) : [-1, 1] \times \mathcal{H}_t \to \mathbb{R}$ is defined as

$$f_t(r_t, h_t) := \int_{\tilde{o}_t, \ldots, \tilde{o}'_1} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1). \tag{5}$$

Therefore, the policy value $\mathcal{V}(\pi)$ can be identified by

$$\mathcal{V}(\pi) = \sum_{t=1}^{T} \int_{r_t} \int_{h_t} r_t \prod_{j=1}^{t} \pi_j(a_j \mid o_j, h_{j-1}) f_t(r_t, h_t). \tag{6}$$

The rationale for expressing the policy value through the sequential integration of bridge functions originates from Lemma B.1 for decomposing $p^\pi(r_t)$ in terms of transition dynamics, reward models and policies, and that $\mathcal{V}(\pi) = \mathbb{E}^\pi\left[\sum_{t=1}^T R_t\right] = \sum_{t=1}^T \int_{\mathcal{R}} r_t p^\pi(r_t)\, dr_t$. Based on the expression of $p^\pi(r_t)$ in Lemma B.1, the challenge lies in learning $\int_{s_t,\ldots,s_1} p(r_t, o_t \mid s_t, a_t) \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j)\, p(s_1)$, which is impeded by the inaccessibility of states directly from observations. Thanks to the introduction of bridge functions and the conditional moment restrictions (Assumption 3.2), we are able to extract the reward and dynamic information from the observable offline dataset. In addition, we establish that $\int_{s_t,\ldots,s_1} p(r_t, o_t \mid s_t, a_t) \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j)\, p(s_1) = \int_{\tilde{o}_t,\ldots,\tilde{o}'_1} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j)\, p(\tilde{o}_1)$. This equation underscores the bridge functions' capacity to encapsulate sufficient reward and dynamic information for policy value identification.

According to Theorem 3.5 and conditional moment restrictions (2)(3), the policy value can be successfully identified under the offline distribution because both bridge functions and conditional moment restrictions rely solely on the observed offline data.

Remark 3.6. It is noteworthy that the marginal distribution of reward $p^\pi(r_t)$ can also be identified according to Theorem 3.5, which is an extra property of our method compared to the model-free methods that directly identify policy values (Bennett & Kallus, 2021; Shi et al., 2022; Miao et al., 2022). As a trade-off, the proposed model-based method requires a sequential integration of bridge functions (4)(5) to identify $p^\pi(r_t)$ for each $t = 1, ..., T$, which might be computationally expensive. Nevertheless, in some RL tasks where maximizing the expected cumulative reward is not the only goal, such as risk-sensitive RL, multi-objective RL, Bayesian RL, the proposed model-based method could also be potentially useful because it is crucial to identify and learn the marginal distribution of reward $p^\pi(r_t)$ in these tasks. We note that the issue of sequential integration involved in learning the marginal distribution can be addressed through the application of Monte Carlo methods, which provide a feasible and efficient way to approximate these integrations.

3.2. Estimation of Bridge Functions

According to Equation (6) presented in Theorem 3.5, to estimate the policy value, it suffices to estimate $f_t(r_t, h_t)$. Further, according to Equation (5), we need to estimate the bridge functions $\{b_R^{[t]}\}_{t=1}^T$, $\{b_D^{[t]}\}_{t=1}^{T-1}$, and $p(\tilde{o}_1)$. We simply assume $p(\tilde{o}_1)$ is known, or we can estimate $p(\tilde{o}_1)$ by its empirical version, denoted as $\widehat{p}(\tilde{o}_1)$. For the two sets of bridge functions, motivated by Singh et al. (2019); Mastouri et al. (2021), we design a two-stage estimation procedure to estimate $\{b_R^{[t]}\}_{t=1}^T$, $\{b_D^{[t]}\}_{t=1}^{T-1}$ according to the conditional moment restrictions, i.e., Equations (2) and (3).

For clarity, we simplify the conditional moment restrictions (2)(3) as

$$\mathbb{E}[b(W, y) \mid X] = p(y \mid X), \tag{7}$$

where we use W to denote (A_t, O_t), y to denote (r_t, o_t) or (o_{t+1}, o_t), and X to denote (H_{t-1}, A_t, O_0), and drop the subscript t. We need to estimate b based on the dataset $(w_n, y_n, x_n)_{n=1}^N$. The estimation procedure can be decomposed into two stages. At the first stage, we learn empirical representations of $p(w \mid x)$ and $p(y \mid x)$ separately. At the second stage, we learn b as a mapping from the representation of $p(w \mid x)$ to the representation of $p(y \mid x)$. RKHS endowed kernel ridge regressions are adopted in this two-stage estimation procedure so that we can obtain a closed-form solution of b. We note that the studied problem in Singh et al. (2019); Mastouri et al. (2021) is $\mathbb{E}[b(W) - Y \mid X] = 0$. We adapt their idea to specifically cater to the problem (7).

We assume that $b(w, y)$ is in the tensor products of $\mathcal{H}_\mathcal{W}$ and $\mathcal{H}_\mathcal{Y}$, i.e. $\mathcal{H}_\mathcal{W} \bigotimes \mathcal{H}_\mathcal{Y}$. That is to say, if $b \in \mathcal{H}_\mathcal{W} \bigotimes \mathcal{H}_\mathcal{Y}$, then $b = \sum_{j=1}^k f_j g_j$ for some $k \in \mathbb{N}$ and such that $f_j \in \mathcal{H}_\mathcal{W}$, $g_j \in \mathcal{H}_\mathcal{Y}$ for all $j \in [k]$. $\mathcal{H}_\mathcal{W}$ and $\mathcal{H}_\mathcal{Y}$ are both set to be RKHSs, which implies that the tensor product $\mathcal{H}_\mathcal{W} \bigotimes \mathcal{H}_\mathcal{Y}$ is also a RKHS. We let $\phi(y)$ denote the canonical feature map of $\mathcal{H}_\mathcal{Y}$. Let $\mu_{W|x} \in \mathcal{H}_\mathcal{W}$ be the conditional mean embedding of $p(W \mid x)$, i.e. $\mu_{W|x} := \int \phi(w) \, \mathrm{d}\, p(w \mid x)$ (Song et al., 2009). Then, we have

$$\mathbb{E}[b(W, y) \mid X = x] = \langle \mu_{W|x} \otimes \phi(y), b \rangle_{\mathcal{H}_\mathcal{W} \otimes \mathcal{H}_\mathcal{Y}} \tag{8}$$

according to Lemma A.1.

According to (7)(8), we design the following risk functional:

$$\mathcal{L}(b) := \int_\mathcal{Y} \mathbb{E}[(\langle \mu_{W|X} \otimes \phi(y), b \rangle_{\mathcal{H}_\mathcal{W} \otimes \mathcal{H}_\mathcal{Y}} - p(y \mid X))^2] dy. \tag{9}$$

The goal is to find $b \in \mathcal{H}_\mathcal{W} \bigotimes \mathcal{H}_\mathcal{Y}$ to minimize $\mathcal{L}(b)$. To achieve this goal, we first learn an empirical estimate of $\mu_{W|x}$ and an empirical estimate of $p(y \mid x)$ at the first stage, denoted as $\widehat{\mu}_{W|x}$ and $\widehat{p}(y \mid x)$ respectively. At the second stage, we learn an estimate of b based on the first stage estimate $\widehat{\mu}_{W|x}$ and $\widehat{p}(y \mid x)$. To alleviate the finite sample bias of the proposed two-stage estimation procedure, we adopt the idea of sample splitting: use N_1 randomly chosen observations in stage 1 and the remaining $N_2 = N - N_1$ observations in stage 2 (Angrist & Krueger, 1995; Singh et al., 2019). We denote the stage 1 observations by $(w_n, y_n, x_n)_{n=1}^{N_1}$ and stage 2 observations by $(w'_n, y'_n, x'_n)_{n=1}^{N_2}$.

Stage 1. From the first sample $(w_n, y_n, x_n)_{n=1}^{N_1}$, we learn the conditional mean embedding of $p(W \mid x)$, i.e., $\widehat{\mu}_{W|x} := \widehat{C}_{W|X}\phi(x)$ where $\widehat{C}_{W|X}$ denotes the conditional mean embedding operator (Song et al., 2013). Specifically, we compute $\widehat{C}_{W|X}$ as a solution to:

$$\begin{aligned} \widehat{C}_{W|X} &= \operatorname*{argmin}_{C \in \mathcal{H}_\Gamma} \widehat{E}(C), \text{ with} \\ \widehat{E}(C) &= \frac{1}{N_1} \sum_{n=1}^{N_1} \|\phi(w_n) - C\phi(x_n)\|_{\mathcal{H}_\mathcal{W}}^2 + \lambda_1 \|C\|_{\mathcal{H}_\Gamma}^2, \end{aligned} \tag{10}$$

where $\mathcal{H}_\Gamma$ is the vector-valued RKHS of operators mapping $\mathcal{H}_\mathcal{X}$ to $\mathcal{H}_\mathcal{W}$. It can be shown that $\widehat{C}_{W|X} = \Phi(W)\left(\mathcal{K}_X + N_1\lambda_1\right)^{-1}\Phi^T(X)$ where $\mathcal{K}_X$ is the $N_1 \times N_1$ kernel matrix and $\Phi(W)$ is a vectors of N_1 columns, with $\phi(w_n)$ in its n th column (Song et al., 2009; Grünewälder et al., 2012; Singh et al., 2019; Mastouri et al., 2021). Consequently, $\widehat{\mu}_{W|x} = \Phi(W)\left(\mathcal{K}_X + N_1\lambda_1\right)^{-1}\mathcal{K}_x$, where K_x is a $N_1 \times 1$ vector with its n-th element denoting $k(x_n, x)$ evaluated at all x_n in the first sample. $\widehat{p}(y \mid x)$ can be learned by any parametric methods like maximum likelihood methods or nonparametric methods like kernel conditional density estimation or generative adversarial networks.

Stage 2. From the second sample $(w'_n, y'_n, x'_n)_{n=1}^{N_2}$, we learn $\widehat{b}$ via empirical risk minimization (ERM):

$$\begin{aligned} \widehat{b} &= \operatorname*{argmin}_{b \in \mathcal{H}_\mathcal{W} \bigotimes \mathcal{H}_\mathcal{Y}} \widehat{\mathcal{L}}(b), \text{ where} \\ \widehat{\mathcal{L}}(b) &= \frac{1}{N_2} \sum_{n=1}^{N_2} \Big(\langle \widehat{\mu}_{W|x'_n} \otimes \phi(y''_n), b \rangle_{\mathcal{H}_\mathcal{W} \otimes \mathcal{H}_\mathcal{Y}} \\ &\quad - \widehat{p}(y''_n \mid x'_n) \Big)^2 + \lambda_2 \|b\|_{\mathcal{H}_\mathcal{W} \otimes \mathcal{H}_\mathcal{Y}}^2, \end{aligned} \tag{11}$$

and $\{y''_n\}_{n=1}^{N_2} \sim_{i.i.d.} \text{unif}(\mathcal{Y})$. The estimator $\widehat{b}$ obtained via the ERM (11) has a closed-form solution $(\widehat{\mathcal{T}}_2 + \lambda_2)^{-1}\widehat{g}_2$ where

$$\begin{aligned} \widehat{\mathcal{T}}_2 &= \frac{1}{N_2} \sum_{n=1}^{N_2} \left[\widehat{\mu}_{W|x'_n} \otimes \phi(y''_n)\right] \otimes \left[\widehat{\mu}_{W|x'_n} \otimes \phi(y''_n)\right], \\ \widehat{g}_2 &= \frac{1}{N_2} \sum_{n=1}^{N_2} \left[\widehat{\mu}_{W|x'_n} \otimes \phi(y''_n)\right] \widehat{p}(y''_n \mid x'_n). \end{aligned} \tag{12}$$

The formula (12) is a direct adaption of Theorem 1 in Mastouri et al. (2021) except that their response variable y is replaced by $\widehat{p}(y \mid x)$ in this work. Indeed,

by the representer theorem (Schölkopf et al., 2001), $\widehat{b}$ can be expressed as a linear combination of feature maps $\widehat{b} = \sum_{i=1}^{N_1}\sum_{i=1}^{N_2} \widehat{c}_{i,j}\phi(w_i)\otimes\phi(y''_j)$ where the coefficients $\widehat{c}_{i,j}$, $i = 1, ..., N_1$, $j = 1, ..., N_2$ are obtained by solving a quadratic minimization problem (11) with respect to $b = \sum_{i=1}^{N_1}\sum_{i=1}^{N_2} c_{i,j}\phi(w_i)\otimes\phi(y''_j)$. See Appendix B.3 and B.5 of Mastouri et al. (2021) for more details on the closed-form solutions to the two-stage estimation procedure.

Next, we adopt the above two-stage estimation procedure to estimate the bridge functions $\{b_R^{[t]}\}_{t=1}^T$, $\{b_D^{[t]}\}_{t=1}^{T-1}$. In particular, we let $\mathcal{W}_t := \mathcal{A}\times\mathcal{O}$, $\mathcal{X}_t := \mathcal{H}_t$, $\mathcal{Y}_t := \mathcal{R}\times\mathcal{O}$, and $\mathcal{Z}_t := \mathcal{O}\times\mathcal{O}$. We define $W_t = (A_t, O_t) \in \mathcal{W}_t$, $X_t = (A_t, H_{t-1}, O_0) \in \mathcal{H}_t$, $Y_t = (R_t, O_t) \in \mathcal{Y}_t$ and $Z_t = (O_{t+1}, O_t) \in \mathcal{Z}_t$. We also define the Hilbert spaces $\mathcal{B}_{R,t} := \mathcal{H}_{\mathcal{W}_t}\bigotimes\mathcal{H}_{\mathcal{Y}_t}$ and $\mathcal{B}_{D,t} := \mathcal{H}_{\mathcal{W}_t}\bigotimes\mathcal{H}_{\mathcal{Z}_t}$ to model the reward-emission bridge function and the dynamic-emission bridge function respectively.

Then we can obtain the empirical conditional mean operator $\widehat{\mu}_{W_t|x_t}$ and the estimates of conditional density functions $\widehat{p}(y_t \mid x_t)$, $\widehat{p}(z_t \mid x_t)$ for $t = 1, ..., T$ according to the introduced state 1 estimation procedure. Subsequently, the estimated bridge functions $\{\widehat{b}_R^{[t]}\}_{t=1}^T$, $\{\widehat{b}_D^{[t]}\}_{t=1}^{T-1}$ can be obtained at the second stage through

$$\widehat{b}_R^{[t]} = \underset{b_{R,t}\in\mathcal{B}_{R,t}}{\operatorname{argmin}}\ \widehat{\mathcal{L}}_R^{[t]}(b_{R,t}), \text{ where}$$
$$\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) = \frac{1}{N_2}\sum_{n=1}^{N_2}\Big(\Big\langle\widehat{\mu}_{W_t|x'_{t,n}}\otimes\phi\left(y''_{t,n}\right), b_{R,t}\Big\rangle_{\mathcal{B}_{R,t}} - \widehat{p}\left(y''_{t,n}\mid x'_{t,n}\right)\Big)^2 + \lambda_2\|b_{R,t}\|^2_{\mathcal{B}_{R,t}}, \tag{13}$$

and $\widehat{b}_D^{[t]} = \underset{b_{D,t}\in\mathcal{B}_{D,t}}{\operatorname{argmin}}\ \widehat{\mathcal{L}}_D^{[t]}(b_{D,t})$, where

$$\widehat{\mathcal{L}}_D^{[t]}(b_{D,t}) = \frac{1}{N_2}\sum_{n=1}^{N_2}\Big(\Big\langle\widehat{\mu}_{W_t|x'_{t,n}}\otimes\phi\left(z''_{t,n}\right), b_{D,t}\Big\rangle_{\mathcal{B}_{D,t}} - \widehat{p}\left(z''_{t,n}\mid x'_{t,n}\right)\Big)^2 + \lambda_2\|b_{D,t}\|^2_{\mathcal{B}_{D,t}}, \tag{14}$$

and $\{y''_{t,n}\}_{n=1}^{N_2}\sim\text{unif}(\mathcal{Y}_t)$, $\{z''_{t,n}\}_{n=1}^{N_2}\sim\text{unif}(\mathcal{Z}_t)$.

3.3. Conservative Policy Optimization within Confidence Regions

Based on (13)(14), we develop two confidence regions for $\mathbf{b}_R := (b_{R,1},\cdots,b_{R,T}) \in \bigotimes_{t=1}^T\mathcal{B}_{R,t}$ and $\mathbf{b}_D := (b_{D,1},\cdots,b_{D,T-1}) \in \bigotimes_{t=1}^{T-1}\mathcal{B}_{D,t}$ as

$$\text{conf}_R(\alpha) = \{\mathbf{b}_R \in \bigotimes_{t=1}^T\mathcal{B}_{R,t} : \widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \widehat{\mathcal{L}}_R^{[t]}(\widehat{b}_R^{[t]}) \le \alpha, \forall t = 1, ..., T\}, \tag{15}$$

Algorithm 1 Conservative model-based policy optimization for POMDPs

Input: Dataset $\mathcal{D}$, regularization parameters λ_1, λ_2, confidence parameters α, β, policy class Π, kernel functions for RKHSs

Estimation of bridge functions: Obtain $\{\widehat{b}_R^{[t]}\}_{t=1}^T$, $\{\widehat{b}_D^{[t]}\}_{t=1}^{T-1}$ by (13)(14)

Construction of confidence regions: Obtain $\text{conf}_R(\alpha)$, $\text{conf}_D(\beta)$ by (15)(16)

Conservative policy optimization:

$$\widehat{\pi} = \arg\max_{\pi\in\Pi}\min_{(\mathbf{b}_R,\mathbf{b}_D)\in\text{conf}_R(\alpha)\times\text{conf}_D(\beta)}\mathcal{V}(\pi,\mathbf{b}_R,\mathbf{b}_D).$$

Return: $\widehat{\pi}$

$$\text{conf}_D(\beta) = \{\mathbf{b}_D \in \bigotimes_{t=1}^{T-1}\mathcal{B}_{D,t} : \widehat{\mathcal{L}}_D^{[t]}(b_{D,t}) - \widehat{\mathcal{L}}_D^{[t]}(\widehat{b}_D^{[t]}) \le \beta, \forall t = 1, ..., T-1\}. \tag{16}$$

where α, β are two constants that will be specified later. Intuitively, these two confidence regions contain all $\mathbf{b}_R \in \bigotimes_{t=1}^T\mathcal{B}_{R,t}$, $\mathbf{b}_D \in \bigotimes_{t=1}^{T-1}\mathcal{B}_{D,t}$ whose risks do not exceed too much than the ones for $\{\widehat{b}_R^{[t]}\}_{t=1}^T$, $\{\widehat{b}_D^{[t]}\}_{t=1}^{T-1}$. They are used to construct a conservative estimate of $\mathcal{V}(\pi)$ under the model-based perspective.

We first use $\mathcal{V}(\pi,\mathbf{b}_R,\mathbf{b}_D)$ to denote the policy value of π by replacing the true bridge functions $\{b_R^{[t]}\}_{t=1}^T$, $\{b_D^{[t]}\}_{t=1}^{T-1}$ with any $\mathbf{b}_R$, $\mathbf{b}_D$ in Theorem 3.5 (policy value identification). We note that $\mathcal{V}(\pi,\{b_R^{[t]}\}_{t=1}^T,\{b_D^{[t]}\}_{t=1}^{T-1})$ is exactly the true policy value $\mathcal{V}(\pi)$. We then define a conservative estimation of $\mathcal{V}(\pi)$ as

$$\underline{\widehat{V}}(\pi) = \min_{(\mathbf{b}_R,\mathbf{b}_D)\in\text{conf}_R(\alpha)\times\text{conf}_D(\beta)}\mathcal{V}(\pi,\mathbf{b}_R,\mathbf{b}_D). \tag{17}$$

Given (17), we propose to choose $\widehat{\pi}$ that maximizes the conservative estimate of policy value $\underline{\widehat{V}}(\pi)$:

$$\widehat{\pi} := \arg\max_{\pi\in\Pi}\underline{\widehat{V}}(\pi). \tag{18}$$

Intuitively, the learned policy $\widehat{\pi}$ defined in (18) aims to maximize the most pessimistic estimator of the policy value within two confidence regions. In Section 4, we provide a a finite-sample upper bound on the suboptimality of $\widehat{\pi}$ under some technical assumptions where only partial coverage of the offline dataset is assumed. We summarize the proposed algorithm in Algorithm 1.

Remark 3.7. We point out that Lu et al. (2022) is the first work to design confidence regions based on empirical risk functionals for bridge functions used in confounded POMDPs, which inspires our construction of confidence regions. Compared to their model-free method, our work

deals with different bridge functions and risk functionals that stem from a model-based perspective. More importantly, the confidence regions used in Lu et al. (2022) depend on policy π, but our proposed ones do not involve π. Such appealing property enjoys both the computational and theoretical advantages. Computationally, any practical algorithm that relies on iterated updates does not need to re-compute the constraint sets at each iteration, therefore greatly reducing the computational cost. Theoretically, the proposed estimation procedure and the construction of confidence regions allow an unrestricted policy class. More details on the comparison to Lu et al. (2022) are provided in the final paragraph of Section 4.

4. Theoretical Results

In this section, we study the theoretical properties of Algorithm 1 under some technical assumptions. We focus on establishing a finite-sample upper bound on the suboptimality of the learned policy $\widehat{\pi}$, i.e., $\mathrm{SubOpt}(\widehat{\pi})$. In particular, such an upper bound will depend on the sample size of the offline data N, the number of stages T, the size of function classes $\{\mathcal{B}_{R,t}\}_{t=1}^{T}$, $\{\mathcal{B}_{D,t}\}_{t=1}^{T-1}$, and the concentrability coefficients $\{C_t^{\pi^*}\}_{t=1}^{T}$, which are defined as

$$C_t^{\pi^*} := \sqrt{\int_{o_t} \mathbb{E}[w_t^{\pi^*}(A_t, H_{t-1}, O_0, o_t)^2]},$$

for $t = 1, \cdots T$. Here $\{w_t^{\pi^*} : \mathcal{A} \times \mathcal{O} \times \mathcal{H}_{t-1} \times \mathcal{O} \to \mathbb{R}\}_{t=1}^{T}$ is a sequence of weight bridge functions that satisfy the following equations for each $t = 1, ...T$

$$\begin{aligned} &\mathbb{E}[w_{\pi^*}^{[t]}(A_t, O_0, H_{t-1}, o_t) \mid S_t, A_t, H_{t-1}] \\ =&\frac{p^{\pi^*}(S_t, H_{t-1})\, \pi_t^*(A_t \mid o_t, H_{t-1})}{p^{\pi^b}(S_t, H_{t-1})\, \pi_t^b(A_t \mid S_t)}. \end{aligned} \tag{19}$$

We assume the existence of such $\{w_t^{\pi^*}\}_{t=1}^{T}$. It can be seen from (19) that the concentrability coefficient $\{C_t^{\pi^*}\}_{t=1}^{T}$ quantifies a distribution mismatch effect between π^b and π^* in a certain sense.

To begin with, we impose the following key assumptions that are used in the theoretical analysis.

Assumption 4.1. The following conditions hold.
(a) (Partial coverage). The concentrability coefficient $C_t^{\pi^*} < \infty, \forall t = 1, ..., T$.
(b) (Consistency of conditional density estimation at stage 1). For any $\delta > 0$, there exist $r_R(\delta, N_1) \to 0$, $r_D(\delta, N_1) \to 0$ as $N_1 \to \infty$ such that for each $t = 1 : T$, with $\mathbb{P}^{\pi^b}$-probability at least $1 - \delta$, the following inequalities hold: $\mathbb{E}_{X_t}[\int_{\mathcal{Y}_t} |\widehat{p}(y_t \mid X_t) - p(y_t \mid X_t)| \,\mathrm{d}\, y_t] \leq r_R(\delta, N_1)$, $\mathbb{E}_{X_t}[\int_{\mathcal{Z}_t} |\widehat{p}(z_t \mid X_t) - p(z_t \mid X_t)| \,\mathrm{d}\, z_t] \leq r_D(\delta, N_1)$.
(c) (Consistency of empirical conditional mean operator at stage 1). For any $\delta > 0$, there exists $r_C(\delta, N_1, c_1) \to 0$, as $N_1 \to \infty$ such that for each $t = 1 : T$ and each $x_t \in \mathcal{X}_t$, with $\mathbb{P}^{\pi^b}$-probability at least $1 - \delta$, it holds that: $\|\widehat{\mu}_{W_t|x_t} - \mu_{W_t|x_t}\|_{\mathcal{H}_{\mathcal{W}_t}} \leq r_C(\delta, N_1, c_1)$. Here c_1 is defined in Assumption D.10 in the appendix.
(d) (Realizability). For each $t = 1 : T$, the RKHSs $\mathcal{B}_{R,t}$, $\mathcal{B}_{D,t}$ contain the solutions $b_R^{[t]}$, $b_D^{[t]}$ to the conditional moment restrictions (2)(3).
(e) (Sizes of $\{\mathcal{B}_{R,t}\}_{t=1}^{T}$, $\{\mathcal{B}_{D,t}\}_{t=1}^{T-1}$). Let $\{\lambda_j^{\downarrow}(K_{\mathcal{F}})\}_{j=1}^{\infty}$ denote the non-increasing eigenvalue sequence of the reproducing kernel $K_{\mathcal{F}}$ for any RKHS $\mathcal{F}$. We assume that $\lambda_j^{\downarrow}(K_{\mathcal{B}_{R,t}}) \asymp j^{-\gamma}$, $\lambda_j^{\downarrow}(K_{\mathcal{B}_{D,t}}) \asymp j^{-\gamma}$ for some $\gamma > 0$.
(f) (Uniform boundness). There exist M_R, $M_D > 0$ such that $\max_{t=1:T} \sup_{b \in \mathcal{B}_{R,t}} \|b\|_\infty \leq M_R$, $\max_{t=1:T-1} \sup_{b \in \mathcal{B}_{D,t}} \|b\|_\infty \leq M_D$, and $\max_{t=1:T} \sup_{p \in \mathcal{P}_{R,t} \cup \mathcal{P}_{D,t}} \|p\|_\infty < \infty$.

Assumption 4.1(a) requires that the offline distribution $\mathbb{P}^{\pi^b}$ can calibrate the distribution induced by the optimal policy π^*. Similar concepts have also been considered in the literature of offline model-free and model-based algorithms for MDPs (Xie et al., 2021; Uehara & Sun, 2021), and model-free algorithms for POMDPs (Lu et al., 2022). This assumption is necessary to ensure the tractability of the problem (Chen & Jiang, 2019). Assumptions 4.1(b)(c) imply that the estimators obtained at stage 1 are sufficiently good to play their roles at stage 2, and thereby guarantee the overall performance of the whole algorithm. In particular, when $\{\mathcal{P}_{R,t}\}_{t=1}^{T}$, $\{\mathcal{P}_{D,t}\}_{t=1}^{T-1}$ are parameterized space and $\widehat{p}$ is obtained by MLE, then $r_R(\delta, N_1)$, $r_D(\delta, N_1)$ usually scale with $\frac{1}{\sqrt{N_1}}$ under some regular conditions on the complexities of $\{\mathcal{P}_{R,t}\}_{t=1}^{T}$, $\{\mathcal{P}_{D,t}\}_{t=1}^{T-1}$ (Geer, 2000). In addition, the convergence rate $r_C(\delta, N_1, c_1)$ for the empirical conditional mean embedding at stage 1 is calibrated by a quantity c_1 that measures the smoothness of $\mu_{W_t|x_t}$, $t = 1, ..., T$ (Singh et al., 2019). Details of results and additional assumptions about realizations of Assumptions 4.1(b)(c) are shown in Appendix D.7. Assumption 4.1(d) requires that the function spaces $\{\mathcal{B}_{R,t}\}_{t=1}^{T}$, $\{\mathcal{B}_{D,t}\}_{t=1}^{T-1}$ are sufficiently large such that there is no model misspecification error when solving the conditional moment restrictions (2)(3). Assumption 4.1(e) requires that RKHSs $\{\mathcal{B}_{R,t}\}_{t=1}^{T}$, $\{\mathcal{B}_{D,t}\}_{t=1}^{T-1}$ enjoy the polynomial eigen-decay rates, which are commonly considered in practice (e.g. Sobolev space). The constants γ quantify the sizes of $\{\mathcal{B}_{R,t}\}_{t=1}^{T}$, $\{\mathcal{B}_{D,t}\}_{t=1}^{T-1}$ in the sense that a larger γ implies smaller $\{\mathcal{B}_{R,t}\}_{t=1}^{T}$, $\{\mathcal{B}_{D,t}\}_{t=1}^{T-1}$. Assumption 4.1(f) is a mild technical condition, which can be easily satisfied.

Next, we present an upper bound on the suboptimality of $\widehat{\pi}$ (18) in the following theorem. All the required lemmas and the complete proof are provided in Appendix C.

Theorem 4.2. *Under Assumptions 3.1, 3.2, 3.3, 4.1, for some constant $c > 0$, by setting $\lambda_1 = N_1^{-\frac{1}{c_1+1}}$, $\lambda_2 =$*

$N_2^{-\frac{\gamma}{\gamma c_2+1}}$, $N_1 = N_2^{\frac{c_1+1}{c_1-1}\frac{\gamma(c_2+1)}{\gamma c_2+1}}$, *and the confidence parameters* α, β *as*

$$\alpha = c\log(T/\delta)M_R N_2^{-\frac{\gamma}{2\gamma+2}}, \beta = c\log(T/\delta)M_D N_2^{-\frac{\gamma}{2\gamma+2}},$$

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

$$\mathrm{SubOpt}(\widehat{\pi}) \lesssim (\sum_{t=1}^{T}(\sqrt{M_R}+(T-t)\sqrt{M_D})C_t^{\pi^*}) \sqrt{\log(T/\delta)}N_2^{-\frac{\gamma}{4\gamma+4}}. \tag{20}$$

Here $C_t^{\pi^*}$, γ, M_R, M_D *are defined in Assumption 4.1(a), (e), (f). Constants* c_1, c_2 *denote a measure of smoothness and are defined in Assumption D.10, D.16 in Appendix D.7.*

According to Theorem 4.2, we have a finite-sample upper bound on the suboptimality of the learned policy $\widehat{\pi}$ in terms of several key parameters. It indicates that the performance of $\widehat{\pi}$ is getting closer to the performance of π^* when the number of samples $N \to \infty$. The constants M_R, M_D denote the uniform upper bounds on the bridge function classes $\{\mathcal{B}_{R,t}\}_{t=1}^T$, $\{\mathcal{B}_{D,t}\}_{t=1}^{T-1}$ respectively. Typically, they do not scale with the number of stages T. Therefore, the upper bound is roughly of the order $(\sum_{t=1}^T (T-t+1)C_t^{\pi^*})\sqrt{\log(T)}N_2^{-\frac{\gamma}{4\gamma+4}}$, where there is a trade-off at the term $(T-t+1)C_t^{\pi^*}$. Intuitively, if t is increasing, then it is harder to require the coverage of trajectory up to stage t, which implies that $C_t^{\pi^*}$ is increasing, and meanwhile the term $T-t+1$ is decreasing. The decay rate γ quantifies the speed of eigen-decay in the RKHS. A rapid decay in eigenvalues (or a large γ) implies that the offline data can be effectively represented in a lower-dimensional subspace of the RKHSs $\{\mathcal{B}_{R,t}\}_{t=1}^T$, $\{\mathcal{B}_{D,t}\}_{t=1}^{T-1}$, suggesting that the kernels capture significant structure in the data with a few dimensions. Therefore, they are often associated with better generalization or faster statistical rate, which is also indicated in the upper bound $O_P(N_2^{\frac{\gamma}{4\gamma+4}})$.

More importantly, the upper bound (20) only relies on the concentrability coefficients $\{C_t^{\pi^*}\}_{t=1}^T$ of the *optimal* policy, which requires that the offline data covers the trajectory generated by the optimal policy π^*. This partial coverage assumption is significantly milder than the restrictive full coverage assumption $\max_{t=1:T}\sup_{\pi\in\Pi} C_t^{\pi} < \infty$ considered in some existing offline methods for confounded POMDPs (Hong et al., 2023).

Remark 4.3. It should be also noted that the upper bound (20) does not involve the size of the policy space $|\Pi|$. This is because the way of estimating bridge functions (13)(14) and constructing confidence sets (15)(16) does not depend on any specific policy π in this work. We take a counterpart result in the model-free method (Lu et al., 2022) as an example to illustrate the advantage of our model-based method. In Theorem 4.4 of Lu et al. (2022), there is a term $\sqrt{\log(|\Pi|)}$ included in the upper bound, which implies that their policy class Π cannot be too large. The effect of this restriction is severe when the optimal policy is not included in the pre-specified limited policy class. Furthermore, under the POMDP settings, we can expect that the dimension of π_t grows with t because of the inclusion of history in π_t. Therefore, it is harder to control $|\Pi|$ in this case. In comparison, the policy space used in our work can be unrestricted and therefore must contain the global optimal policy as long as its concentrability coefficient is finite. This property is especially meaningful when the policy update rule does not need explicit policy parameterization (see e.g. Lan (2022)).

5. Discussion

We propose a model-based offline RL method for confounded POMDPs. Under some mild conditions, we establish a finite-sample upper bound on the performance of the learned policy under the partial coverage assumption from a model-based perspective.

We present some discussions and limitations in this section. First, it would be intriguing to design a practical algorithm with further empirical evaluation to demonstrate the practical effectiveness of the proposed method. In particular, since RKHS can be employed for modeling bridge functions, the bridge functions can be expressed as linear combinations of many feature functions, making the ERM a quadratic function with respect to the coefficients associated with the bridge functions. As a result, the estimators of the bridge functions will have closed forms, making them computationally tractable and applicable to subsequent tasks. To perform conservative policy optimization, the idea of an existing work that designed a practical pessimistic model-based algorithm in standard MDP contexts (Rigter et al., 2022) could be potentially adapted to our confounded POMDP settings. Second, in this paper, we focus on the case when the bridge functions are realizable (Assumption 4.1(d)), the estimated conditional density functions at stage 1 are consistent (Assumption 4.1(b)), and the empirical conditional mean operator at stage 1 is consistent (Assumption 4.1(c)). In other words, all the required function spaces for the bridge functions, conditional density functions, and conditional mean operators are sufficiently large so that there is no approximation error occurring in this work. It would be interesting to relax these assumptions and allow for approximation error. Techniques like balancing the estimation error and approximation error could potentially be applied to broaden the applicability of the method. Lastly, it would be also meaningful to apply the established model-based identification in some RL tasks where maximizing the expected cumulative reward is not the only goal, but also controlling the risk or optimizing some other objectives.

Acknowledgements

The authors wish to thank anonymous reviewers for their valuable feedback on an early version of this paper. The work of Xu was supported by NSF 1918854, NSF 1940107, and NIH R01MH128085.

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 which we feel must be specifically highlighted here.

References

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107, 2020.

Albright, S. C. Structural results for partially observable markov decision processes. *Operations Research*, 27(5): 1041–1053, 1979.

Angrist, J. D. and Krueger, A. B. Split-sample instrumental variables estimates of the return to schooling. *Journal of Business & Economic Statistics*, 13(2):225–235, 1995.

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

Bravo, R. Z. B., Leiras, A., and Cyrino Oliveira, F. L. The use of uav s in humanitarian relief: an application of pomdp-based methodology for finding victims. *Production and Operations Management*, 28(2):421–440, 2019.

Burago, D., De Rougemont, M., and Slissenko, A. On the complexity of partially observed markov decision processes. *Theoretical Computer Science*, 157(2):161–183, 1996.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

Cassandra, A. R. A survey of pomdp applications. In *Working notes of AAAI 1998 fall symposium on planning with partially observable Markov decision processes*, volume 1724, 1998.

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

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

Fu, Z., Qi, Z., Wang, Z., Yang, Z., Xu, Y., and Kosorok, M. R. Offline reinforcement learning with instrumental variables in confounded markov decision processes. *arXiv preprint arXiv:2209.08666*, 2022.

Geer, S. A. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., and Celi, L. A. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019.

Grünewälder, S., Lever, G., Baldassarre, L., Patterson, S., Gretton, A., and Pontil, M. Conditional mean embeddings as regressors-supplementary. *arXiv preprint arXiv:1205.4656*, 2012.

Grunewalder, S., Lever, G., Baldassarre, L., Pontil, M., and Gretton, A. Modelling transition dynamics in mdps with rkhs embeddings. *arXiv preprint arXiv:1206.4655*, 2012.

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

Hu, Y. and Shiu, J.-L. Nonparametric identification using instrumental variables: sufficient conditions for completeness. *Econometric Theory*, 34(3):659–693, 2018.

Jaakkola, T., Singh, S., and Jordan, M. Reinforcement learning algorithm for partially observable markov decision problems. *Advances in neural information processing systems*, 7, 1994.

Kress, R., Maz'ya, V., and Kozlov, V. *Linear integral equations*, volume 82. Springer, 1989.

Lan, G. Policy optimization over general state and action spaces. *arXiv preprint arXiv:2211.16715*, 2022.

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.

Lu, M., Min, Y., Wang, Z., and Yang, Z. Pessimism in the face of confounders: Provably efficient offline reinforcement learning in partially observable markov decision processes. *arXiv preprint arXiv:2205.13589*, 2022.

Mastouri, A., Zhu, Y., Gultchin, L., Korba, A., Silva, R., Kusner, M., Gretton, A., and Muandet, K. Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International conference on machine learning*, pp. 7512–7523. PMLR, 2021.

Miao, R., Qi, Z., and Zhang, X. Off-policy evaluation for episodic partially observable markov decision processes under non-parametric models. *arXiv preprint arXiv:2209.10064*, 2022.

Miao, W., Shi, X., and Tchetgen, E. T. A confounding bridge approach for double negative control inference on causal effects. *arXiv preprint arXiv:1808.04945*, 2018.

Monahan, G. E. State of the art—a survey of partially observable markov decision processes: theory, models, and algorithms. *Management science*, 28(1):1–16, 1982.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

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

Newey, W. K. and Powell, J. L. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5): 1565–1578, 2003.

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

Rigter, M., Lacerda, B., and Hawes, N. Rambo-rl: Robust adversarial model-based offline reinforcement learning. *Advances in neural information processing systems*, 35: 16082–16097, 2022.

Sawaki, K. and Ichikawa, A. Optimal control for partially observable markov decision processes over an infinite horizon. *Journal of the Operations Research Society of Japan*, 21(1):1–16, 1978.

Schölkopf, B., Herbrich, R., and Smola, A. J. A generalized representer theorem. In *International conference on computational learning theory*, pp. 416–426. Springer, 2001.

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.

Singh, R., Sahani, M., and Gretton, A. Kernel instrumental variable regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Singh, S. P., Jaakkola, T., and Jordan, M. I. Learning without state-estimation in partially observable markovian decision processes. In *Machine Learning Proceedings 1994*, pp. 284–292. Elsevier, 1994.

Song, L., Huang, J., Smola, A., and Fukumizu, K. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 961–968, 2009.

Song, L., Fukumizu, K., and Gretton, A. Kernel embeddings of conditional distributions: A unified kernel framework for nonparametric inference in graphical models. *IEEE Signal Processing Magazine*, 30(4):98–111, 2013.

Sriperumbudur, B. K., Fukumizu, K., and Lanckriet, G. R. Universality, characteristic kernels and rkhs embedding of measures. *Journal of Machine Learning Research*, 12 (7), 2011.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

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

Szabó, Z., Sriperumbudur, B. K., Póczos, B., and Gretton, A. Learning theory for distribution regression. *The Journal of Machine Learning Research*, 17(1):5272–5311, 2016.

Tchetgen, E. J. T., 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 10276–10283, 2020.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Vito, E. D. and Caponnetto, A. Risk bounds for regularized least-squares algorithm with operator-valued kernels. 2005.

Vlassis, N., Littman, M. L., and Barber, D. On the computational complexity of stochastic controller optimization in pomdps. *ACM Transactions on Computation Theory (TOCT)*, 4(4):1–8, 2012.

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

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Young, S., Gašić, M., Thomson, B., and Williams, J. D. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 101(5):1160–1179, 2013.

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

A. Definitions and Auxiliary Lemmas

Lemma A.1. *Let $\mu_{W|x}$ be the conditional mean embedding of $p(W \mid x)$, then it holds that*

$$\mathbb{E}[b(W,y) \mid X = x] = \langle b, \mu_{W|x} \otimes \phi(y) \rangle_{\mathcal{H}_\mathcal{W} \otimes \mathcal{H}_\mathcal{Y}} \tag{21}$$

where $\phi(y)$ denotes the canonical feature map of $\mathcal{H}_\mathcal{Y}$.

Definition A.2 (Population risk functional)**.** For any generic $b_{R,t}$ and $t = 1, ..., T$, the population risk functional $\mathcal{L}_R^{[t]}(b_{R,t})$ is defined as

$$\mathcal{L}_R^{[t]}(b_{R,t}) = \frac{1}{\mathrm{vol}(\mathcal{R} \times \mathcal{O})} \int_{\mathcal{R}\times\mathcal{O}} \mathbb{E}\left[(p(r_t, o_t \mid A_t, H_{t-1}, O_0) - \mathbb{E}[b_{R,t}(A_t, O_t, r_t, o_t) \mid A_t, H_{t-1}, O_0])^2 \right] dr_t do_t. \tag{22}$$

Similarly, for any generic $b_{D,t}$ and $t = 1, ..., T-1$, the population risk functional $\mathcal{L}_D^{[t]}(b_{D,t})$ is defined as

$$\mathcal{L}_D^{[t]}(b_{D,t}) = \frac{1}{\mathrm{vol}(\mathcal{O} \times \mathcal{O})} \int_{\mathcal{O}\times\mathcal{O}} \mathbb{E}\left[(p(o_{t+1}, o_t \mid A_t, H_{t-1}, O_0) - \mathbb{E}[b_{D,t}(A_t, O_t, o_{t+1}, o_t) \mid A_t, H_{t-1}, O_0])^2 \right] do_{t+1} do_t. \tag{23}$$

Definition A.3 (Concentrability coefficient)**.** For each $\pi \in \Pi$, the concentrability coefficient at each stage $t = 1, ..., T$ is defined as

$$C_t^\pi = \sqrt{\int_{o_t} \mathbb{E}[w_t^\pi(A_t, H_{t-1}, O_0, o_t)^2]} \tag{24}$$

where w_t^π is defined in Assumption 3.2.

Definition A.4 (Star convex hull of $\mathcal{H}$)**.** For a function class $\mathcal{H}$, we define $\mathrm{star}(\mathcal{H}) := \{rh : h \in \mathcal{H}, r \in [-1, 1]\}$.

Lemma A.5 (Lemma 14 of (Foster & Syrgkanis, 2023))**.** *Consider a function class $\mathcal{F}$, with $\sup_{f\in\mathcal{F}} \|f\|_\infty \le 1$, and pick any $f^\star \in \mathcal{F}$. Let $\delta_n^2 \ge \frac{4d\log(41\log(2c_2 n))}{c_2 n}$ be any solution to the inequalities:*

$$\mathcal{R}\left(\delta, \mathrm{star}\left(\mathcal{F} - f^\star\right)\right) \le \delta^2.$$

Moreover, assume that the loss ℓ is L-Lipschitz in its first argument with respect to the ℓ_2 norm. Then for some universal constants c_5, c_6, with probability $1 - c_5 \exp\left(c_6 n \delta_n^2\right)$,

$$\begin{aligned} &\left| \left(\widehat{\mathbb{E}}_n[\ell(f(x), y)] - \widehat{\mathbb{E}}_n\left[\ell\left(f^\star(x), y\right)\right] \right) - \left(\mathbb{E}[\ell(f(x), y)] - \mathbb{E}\left[\ell\left(f^\star(x), y\right)\right]\right) \right| \\ &\quad \le 18L\delta_n \left(\|f - f^\star\|_2 + \delta_n\right), \ \forall f \in \mathcal{F}. \end{aligned} \tag{25}$$

B. Proof of Theorem 3.5

In this section, we present a complete proof of the identification results summarized in Theorem 3.5. In the first part, we show that under Assumptions 3.1, 3.2, 3.3, we have a sequence of conditional moment restrictions that are conditioned on the unobserved (S_t, A_t). In the second part, we derive the identification results based on the first part and conclude the proof.

Part I. According to Assumption 3.2, the following two equations hold almost surely with respect to $\mathbb{P}^{\pi^b}$:

$$\mathbb{E}[b_R^{[t]}(A_t, O_t, r_t, o_t) \mid H_{t-1}, A_t, O_0] = p(r_t, o_t \mid H_{t-1}, A_t, O_0), \tag{26}$$

$$\mathbb{E}[b_D^{[t]}(A_t, O_t, o_{t+1}, o_t) | H_{t-1}, A_t, O_0] = p(o_{t+1}, o_t \mid H_{t-1}, A_t, O_0). \tag{27}$$

In the rest of this part, we will show that equations (26)(27) also hold when the projected space is replaced by the one generated by the unobserved (S_t, A_t), under Assumption 3.3.

We first analyze the equation (26). The LHS of (26) can be written as

$$\begin{aligned} &\mathbb{E}[b_R^{[t]}(A_t, O_t, r_t, o_t) \mid H_{t-1}, A_t, O_0] \\ =&\mathbb{E}[\mathbb{E}[b_R^{[t]}(A_t, O_t, r_t, o_t) \mid S_t, H_{t-1}, A_t, O_0] \mid H_{t-1}, A_t, O_0] \\ =&\mathbb{E}[\mathbb{E}[b_R^{[t]}(A_t, O_t, r_t, o_t) \mid S_t, H_{t-1}, A_t] \mid H_{t-1}, A_t, O_0] \\ =&\mathbb{E}[\mathbb{E}[b_R^{[t]}(A_t, O_t, r_t, o_t) \mid S_t, A_t] \mid H_{t-1}, A_t, O_0]. \end{aligned} \tag{28}$$

The first equality comes from the law of total expectation. The second equality comes from Assumption 3.1: $O_t \perp\!\!\!\perp O_0 \mid S_t, A_t, H_{t-1}$. The last equality is due to $O_t \perp\!\!\!\perp H_{t-1} \mid S_t, A_t$.

The RHS of (26) can be written as

$$\begin{aligned} & p(r_t, o_t \mid H_{t-1}, A_t, O_0) \\ =& \mathbb{E}[p(r_t, o_t \mid S_t, H_{t-1}, A_t, O_0) \mid H_{t-1}, A_t, O_0] \\ =& \mathbb{E}[p(r_t, o_t \mid S_t, A_t) \mid H_{t-1}, A_t, O_0] \end{aligned} \tag{29}$$

where the last equality is due to $(R_t, O_t) \perp\!\!\!\perp (H_{t-1}, O_0) \mid S_t, A_t$ based on the data generating processes. More specifically, given (S_t, A_t) and under the offline distribution $\mathbb{P}^{\pi^b}$, O_t depends on S_t through the observation emission kernel, and R_t depends on (S_t, A_t) through the reward kernel. Therefore, (R_t, O_t) are conditional independent of (H_{t-1}, O_0) given the state-action pair (S_t, A_t).

By combining equations (28)(29), we have

$$\mathbb{E}[\mathbb{E}[b_R^{[t]}(A_t, O_t, r_t, o_t) \mid S_t, A_t] \mid H_{t-1}, A_t, O_0] = \mathbb{E}[p(r_t, o_t \mid S_t, A_t) \mid H_{t-1}, A_t, O_0], \tag{30}$$

which means

$$\mathbb{E}[\mathbb{E}[b_R^{[t]}(A_t, O_t, r_t, o_t) \mid S_t, A_t] - p(r_t, o_t \mid S_t, A_t) \mid H_{t-1}, A_t, O_0] = 0. \tag{31}$$

Then, the combination of Assumption 3.3 and equation (31) implies that

$$\mathbb{E}[b_R^{[t]}(A_t, O_t, r_t, o_t) \mid S_t, A_t] = p(r_t, o_t \mid S_t, A_t) \tag{32}$$

for all $(r_t, o_t) \in \mathcal{R} \times \mathcal{O}$, $\mathbb{P}^{\pi^b}$-a.s. (S_t, A_t), and for all $t = 1, ..., T$.

Next, we use almost the same arguments to analyze the equation (27). The LHS of (27) can be written as

$$\begin{aligned} & \mathbb{E}[b_D^{[t]}(A_t, O_t, o_{t+1}, o_t) \mid H_{t-1}, A_t, O_0] \\ =& \mathbb{E}[\mathbb{E}[b_D^{[t]}(A_t, O_t, o_{t+1}, o_t) \mid S_t, H_{t-1}, A_t, O_0] \mid H_{t-1}, A_t, O_0] \\ =& \mathbb{E}[\mathbb{E}[b_D^{[t]}(A_t, O_t, o_{t+1}, o_t) \mid S_t, H_{t-1}, A_t] \mid H_{t-1}, A_t, O_0] \\ =& \mathbb{E}[\mathbb{E}[b_D^{[t]}(A_t, O_t, o_{t+1}, o_t) \mid S_t, A_t] \mid H_{t-1}, A_t, O_0] \end{aligned} \tag{33}$$

Similarly, the first equality comes from the law of total expectation. The second equality comes from Assumption 3.1: $O_t \perp\!\!\!\perp O_0 \mid S_t, A_t, H_{t-1}$. The last equality is due to $O_t \perp\!\!\!\perp H_{t-1} \mid S_t, A_t$.

The RHS of (27) can be written as

$$\begin{aligned} & p(o_{t+1}, o_t \mid H_{t-1}, A_t, O_0) \\ =& \mathbb{E}[p(o_{t+1}, o_t \mid S_t, H_{t-1}, A_t, O_0) \mid H_{t-1}, A_t, O_0] \\ =& \mathbb{E}[p(o_{t+1}, o_t \mid S_t, A_t) \mid H_{t-1}, A_t, O_0] \end{aligned} \tag{34}$$

where the last equality is due to $(O_{t+1}, O_t) \perp\!\!\!\perp (H_{t-1}, O_0) \mid S_t, A_t$ based on the data generating processes. More specifically, given (S_t, A_t) and under the offline distribution $\mathbb{P}^{\pi^b}$, O_t depends on S_t through the observation emission kernel, and O_{t+1} depends on (S_t, A_t) through the observation emission kernel at time $t+1$ as well as the transition kernel at time t, i.e. $O_{t+1} \sim \mathcal{E}(S_{t+1})$, $S_{t+1} \sim P_t(\cdot \mid S_t, A_t)$. Therefore, (O_{t+1}, O_t) are conditional independent of (H_{t-1}, O_0) given the state-action pair (S_t, A_t).

By combining equations (33)(34), we have

$$\mathbb{E}[\mathbb{E}[b_D^{[t]}(A_t, O_t, o_{t+1}, o_t) \mid S_t, A_t] \mid H_{t-1}, A_t, O_0] = \mathbb{E}[p(o_{t+1}, o_t \mid S_t, A_t) \mid H_{t-1}, A_t, O_0], \tag{35}$$

which means

$$\mathbb{E}[\mathbb{E}[b_D^{[t]}(A_t, O_t, o_{t+1}, o_t) \mid S_t, A_t] - p(o_{t+1}, o_t \mid S_t, A_t) \mid H_{t-1}, A_t, O_0] = 0. \tag{36}$$

Then, the combination of Assumption 3.3 and equation (36) implies that

$$\mathbb{E}[b_D^{[t]}(A_t, O_t, o_{t+1}, o_t) \mid S_t, A_t] = p(o_{t+1}, o_t \mid S_t, A_t) \tag{37}$$

for all $(o_{t+1}, o_t) \in \mathcal{O} \times \mathcal{O}$, $\mathbb{P}^{\pi^b}$-a.s. (S_t, A_t), and for all $t = 1, ..., T-1$.

In summary, we have shown in part I that under Assumptions 3.1, 3.2, 3.3, the following equations hold for all $(r_t, o_t, o_{t+1}) \in \mathcal{R} \times \mathcal{O} \times \mathcal{O}$ and $\mathbb{P}^{\pi^b}$-a.s. (S_t, A_t).

$$\mathbb{E}[b_R^{[t]}(A_t, O_t, r_t, o_t) \mid S_t, A_t] = p(r_t, o_t \mid S_t, A_t), \ \forall t = 1, ..., T \tag{38}$$

$$\mathbb{E}[b_D^{[t]}(A_t, O_t, o_{t+1}, o_t) \mid S_t, A_t] = p(o_{t+1}, o_t \mid S_t, A_t), \ \forall t = 1, ..., T-1. \tag{39}$$

Part II.

By the definition of the policy value

$$\mathcal{V}(\pi) = \mathbb{E}^{\pi}[\sum_{t=1}^{T} R_t] = \sum_{t=1}^{T} \int_{\mathcal{R}} r_t p^{\pi}(r_t) dr_t,$$

it remains to identify the marginal distribution of the reward R_t induced by the policy π, i.e. $p^{\pi}(r_t)$. In the following lemma, we express $p^{\pi}(r_t)$ in terms of the combination of policy functions π_t, reward-emission models $p(r_t, o_t \mid s_t, a_t)$, and dynamic-emission models $p(s_{t+1}, o_t \mid s_t, a_t)$.

Lemma B.1. *For each $1 \leq t \leq T$, we have*

$$p^{\pi}(r_t) = \sum_{a_t, a_{t-1}, ..., a_1} \int_{o_t, o_{t-1}, ..., o_1} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, a_{j-1}, o_{j-1}, ...) \int_{s_t, s_{t-1}, ..., s_1} p(r_t, o_t \mid s_t, a_t) \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1). \tag{40}$$

Lemma B.1 has also been used in Tennenholtz et al. (2020) which focuses on the tabular settings. We generalize it to the continuous settings by simply extending summation to integration. The proof of Lemma B.1 is directed adapted from Tennenholtz et al. (2020). For completeness, we provide a proof in Appendix D.2.

According to Lemma B.1, in order to identify $p^{\pi}(r_t)$, it suffices to identify the following function

$$f_t(r_t, h_t) := \int_{s_t, s_{t-1}, ..., s_1} p(r_t, o_t \mid s_t, a_t) \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1). \tag{41}$$

which encodes both the information of the reward model and the dynamic model at each step t under the offline distribution. According to Lemma B.1 and the definition of f_t, it can be seen easily that

$$p^{\pi}(r_t) = \sum_{a_t, a_{t-1}, ..., a_1} \int_{o_t, o_{t-1}, ..., o_1} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, h_{j-1}) f_t(r_t, h_t). \tag{42}$$

In the rest of part II, we present a novel analysis on $f_t(r_t, h_t)$, proving that it can be identified under Assumptions 3.1, 3.2, 3.3 under the general function approximation settings. In particular, we show that it can be expressed as the form of sequential integration of bridge functions.

We focus on

$$f_t(r_t, h_t) = \int_{s_t, s_{t-1}, ..., s_1} p(r_t, o_t \mid s_t, a_t) \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1). \tag{43}$$

We first look at the term $p(r_t, o_t \mid s_t, a_t)$ in (43). According to the results (38) shown in part I, we have

$$\mathbb{E}[b_R^{[t]}(A_t, O_t, r_t, o_t) \mid S_t, A_t] = p(r_t, o_t \mid S_t, A_t). \tag{44}$$

By plugging (44) into (43), it holds that

$$
\begin{aligned}
& f_t(r_t, h_t) \\
&= \int_{s_t, s_{t-1}, \ldots, s_1} p(r_t, o_t \mid s_t, a_t) \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
&= \int_{s_t, s_{t-1}, \ldots, s_1} \mathbb{E}[b_R^{[t]}(A_t, O_t, r_t, o_t) \mid S_t = s_t, A_t = a_t] \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \text{ by (44)} \\
&= \int_{s_t, s_{t-1}, \ldots, s_1} \mathbb{E}[b_R^{[t]}(a_t, O_t, r_t, o_t) \mid S_t = s_t, A_t = a_t] \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
&= \int_{s_t, s_{t-1}, \ldots, s_1} \mathbb{E}[b_R^{[t]}(a_t, O_t, r_t, o_t) \mid S_t = s_t] \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \text{ by } O_t \perp\!\!\!\perp A_t \mid S_t \\
&= \int_{s_t, s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) p(\tilde{o}_t \mid s_t) \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
&= \int_{s_t, s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) p(\tilde{o}_t \mid s_t) p(s_t, o_{t-1} \mid s_{t-1}, a_{t-1}) \prod_{j=1}^{t-2} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
&= \int_{s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \Big(\int_{s_t} p(\tilde{o}_t \mid s_t) p(s_t, o_{t-1} \mid s_{t-1}, a_{t-1}) \Big) \prod_{j=1}^{t-2} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
&= \int_{s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \Big(\int_{s_t} p(\tilde{o}_t \mid s_t, o_{t-1}, s_{t-1}, a_{t-1}) p(s_t, o_{t-1} \mid s_{t-1}, a_{t-1}) \Big) \prod_{j=1}^{t-2} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
&= \int_{s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \Big(\int_{s_t} p(\tilde{o}_t, s_t, o_{t-1} \mid s_{t-1}, a_{t-1}) \Big) \prod_{j=1}^{t-2} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
&= \int_{s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) p(\tilde{o}_t, o_{t-1} \mid s_{t-1}, a_{t-1}) \prod_{j=1}^{t-2} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \qquad (45)
\end{aligned}
$$

Next, we look at the term $p(\tilde{o}_t, o_{t-1} \mid s_{t-1}, a_{t-1})$ in the last equality of (45). According to the results (39) shown in part I, we have

$$
\mathbb{E}[b_D^{[t-1]}(A_{t-1}, O_{t-1}, \tilde{o}_t, o_{t-1}) \mid S_{t-1}, A_{t-1}] = p(\tilde{o}_t, o_{t-1} \mid S_{t-1}, A_{t-1}). \qquad (46)
$$

By plugging (46) into the last equality of (45), we have

$$
\begin{aligned}
& f_t(r_t, h_t) \\
&= \int_{s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) p(\tilde{o}_t, o_{t-1} \mid s_{t-1}, a_{t-1}) \prod_{j=1}^{t-2} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \text{ by (45)} \\
&= \int_{s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \mathbb{E}[b_D^{[t-1]}(A_{t-1}, O_{t-1}, \tilde{o}_t, o_{t-1}) \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}] \\
& \quad \prod_{j=1}^{t-2} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \text{ by (46)} \\
&= \int_{s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \mathbb{E}[b_D^{[t-1]}(a_{t-1}, O_{t-1}, \tilde{o}_t, o_{t-1}) \mid S_{t-1} = s_{t-1}, A_{t-1} = a_{t-1}]
\end{aligned}
$$

$$
\begin{aligned}
&\prod_{j=1}^{t-2} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
=&\int_{s_{t-1},\dots,s_1} \int_{\tilde{o}_t} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \mathbb{E}[b_D^{[t-1]}(a_{t-1}, O_{t-1}, \tilde{o}_t, o_{t-1}) \mid S_{t-1} = s_{t-1}] \\
&\prod_{j=1}^{t-2} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \text{ by } O_{t-1} \perp\!\!\!\perp A_{t-1} \mid S_{t-1} \\
=&\int_{s_{t-1},\dots,s_1} \int_{\tilde{o}_t} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \int_{\tilde{o}_{t-1}} b_D^{[t-1]}(a_{t-1}, \tilde{o}_{t-1}, \tilde{o}_t, o_{t-1}) p(\tilde{o}_{t-1} \mid s_{t-1}) \prod_{j=1}^{t-2} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
=&\int_{s_{t-1},\dots,s_1} \int_{\tilde{o}_t} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \int_{\tilde{o}_{t-1}} b_D^{[t-1]}(a_{t-1}, \tilde{o}_{t-1}, \tilde{o}_t, o_{t-1}) p(\tilde{o}_{t-1} \mid s_{t-1}) p(s_{t-1}, o_{t-2} \mid s_{t-2}, a_{t-2}) \\
&\prod_{j=1}^{t-3} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
=&\int_{s_{t-2},\dots,s_1} \int_{\tilde{o}_t, \tilde{o}_{t-1}} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) b_D^{[t-1]}(a_{t-1}, \tilde{o}_{t-1}, \tilde{o}_t, o_{t-1}) (\int_{s_{t-1}} p(\tilde{o}_{t-1} \mid s_{t-1}) p(s_{t-1}, o_{t-2} \mid s_{t-2}, a_{t-2})) \\
&\prod_{j=1}^{t-3} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
=&\int_{s_{t-2},\dots,s_1} \int_{\tilde{o}_t, \tilde{o}_{t-1}} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) b_D^{[t-1]}(a_{t-1}, \tilde{o}_{t-1}, \tilde{o}_t, o_{t-1}) (\int_{s_{t-1}} p(\tilde{o}_{t-1}, s_{t-1}, o_{t-2} \mid s_{t-2}, a_{t-2})) \\
&\prod_{j=1}^{t-3} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
=&\int_{s_{t-2},\dots,s_1} \int_{\tilde{o}_t, \tilde{o}_{t-1}} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) b_D^{[t-1]}(a_{t-1}, \tilde{o}_{t-1}, \tilde{o}_t, o_{t-1}) p(\tilde{o}_{t-1}, o_{t-2} \mid s_{t-2}, a_{t-2}) \prod_{j=1}^{t-3} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1).
\end{aligned} \tag{47}
$$

Based on the derivations in (45)(47), we have shown the ways to tackle the reward-emission model $p(r_t, o_t \mid s_t, a_t)$ at time t and the dynamic-emission model $p(\tilde{o}_t, o_{t-1} \mid s_{t-1}, a_{t-1})$ at the time $t-1$. By repeating the procedure of tackling $p(\tilde{o}_j, o_{j-1} \mid s_{j-1}, a_{j-1})$ at the time $j = 1, ..., t-2$ along with

$$
\mathbb{E}[b_D^{[j-1]}(A_{j-1}, O_{j-1}, \tilde{o}_j, o_{j-1}) \mid S_{j-1}, A_{j-1}] = p(\tilde{o}_j, o_{j-1} \mid S_{j-1}, A_{j-1}),\ \forall j = 1, ..., t-1, \tag{48}
$$

we can write the last equality of (47) as

$$
\begin{aligned}
&f_t(r_t, h_t) \\
=&\int_{s_{t-2},\dots,s_1} \int_{\tilde{o}_t, \tilde{o}_{t-1}} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) b_D^{[t-1]}(a_{t-1}, \tilde{o}_{t-1}, \tilde{o}_t, o_{t-1}) p(\tilde{o}_{t-1}, o_{t-2} \mid s_{t-2}, a_{t-2}) \\
&\prod_{j=1}^{t-3} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \text{ by (47)} \\
=&\int_{s_{t-3},\dots,s_1} \int_{\tilde{o}_t, \tilde{o}_{t-1}, \tilde{o}_{t-2}} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) b_D^{[t-1]}(a_{t-1}, \tilde{o}_{t-1}, \tilde{o}_t, o_{t-1}) b_D^{[t-2]}(a_{t-2}, \tilde{o}_{t-2}, \tilde{o}_{t-1}, o_{t-2}) \\
&p(\tilde{o}_{t-2}, o_{t-3} \mid s_{t-3}, a_{t-3}) \prod_{j=1}^{t-4} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1)
\end{aligned}
$$

$$
\begin{aligned}
&= \cdots \\
&= \int_{s_1} \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_2} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=2}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_2, o_1 \mid s_1, a_1) p(s_1) \\
&= \int_{s_1} \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_2} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=2}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) \mathbb{E}[b_D^{[1]}(A_1, O_1, \tilde{o}_2, o_1) \mid S_1 = s_1, A_1 = a_1] p(s_1) \\
&= \int_{s_1} \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_2} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=2}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) \mathbb{E}[b_D^{[1]}(a_1, O_1, \tilde{o}_2, o_1) \mid S_1 = s_1, A_1 = a_1] p(s_1) \\
&= \int_{s_1} \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_2} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=2}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) \mathbb{E}[b_D^{[1]}(a_1, O_1, \tilde{o}_2, o_1) \mid S_1 = s_1] p(s_1) \text{ by } O_1 \perp\!\!\!\perp A_1 \mid S_1 \\
&= \int_{s_1} \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_2} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=2}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) \int_{\tilde{o}_1} b_D^{[1]}(a_1, \tilde{o}_1, \tilde{o}_2, o_1) p(\tilde{o}_1 \mid s_1) p(s_1) \\
&= \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_2} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=2}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) \int_{\tilde{o}_1} b_D^{[1]}(a_1, \tilde{o}_1, \tilde{o}_2, o_1) p(\tilde{o}_1) \text{ by integrating out } s_1 \\
&= \int_{\tilde{o}_t, \ldots \tilde{o}_1} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1).
\end{aligned} \tag{49}
$$

Consequently, for each t, $f_t(r_t, h_t)$ can be identified as

$$
f_t(r_t, h_t) = \int_{\tilde{o}_t, \ldots \tilde{o}_1} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1), \tag{50}
$$

which implies the identification of $p^\pi(r_t)$ for each $t = 1, ..., T$:

$$
\begin{aligned}
& p^\pi(r_t) \\
&= \sum_{a_t, \ldots, a_1} \int_{o_t, \ldots, o_1} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, a_{j-1}, o_{j-1}, ...) f_t(r_t, h_t) \\
&= \sum_{a_t, \ldots, a_1} \int_{o_t, \ldots, o_1} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, a_{j-1}, o_{j-1}, ...) \int_{\tilde{o}_t, \ldots \tilde{o}_1} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1).
\end{aligned} \tag{51}
$$

The policy value $\mathcal{V}(\pi)$ is then identified as

$$
\begin{aligned}
& \mathcal{V}(\pi) \\
&= \mathbb{E}^\pi[\sum_{t=1}^{T} R_t] \\
&= \sum_{t=1}^{T} \int_{r_t} r_t p^\pi(r_t) \\
&= \sum_{t=1}^{T} \int_{r_t} r_t \sum_{a_t, \ldots, a_1} \int_{o_t, \ldots, o_1} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, a_{j-1}, o_{j-1}, ...) \int_{\tilde{o}_t, \ldots \tilde{o}_1} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1).
\end{aligned} \tag{52}
$$

The proof of Theorem 3.5 is completed.

C. Proof of Theorem 4.2

In this section, we provide a detailed proof of Theorem 4.2, which provides an upper bound on the suboptimality of $\widehat{\pi}$.

We rely on the following three useful lemmas. The first lemma is quantifies the performance difference between the true bridge function and the function in the bridge function class.

Lemma C.1 (Error decomposition)**.** *Under Assumptions 3.1, 3.2, 3.3, D.3, for each* $\pi \in \Pi$, $\mathbf{b}_R \in \bigotimes_{t=1}^{T} \mathcal{H}_{\mathcal{W}_t} \bigotimes \mathcal{H}_{\mathcal{Y}_t}$, $\mathbf{b}_D \in \bigotimes_{t=1}^{T-1} \mathcal{H}_{\mathcal{W}_t} \bigotimes \mathcal{H}_{\mathcal{Z}_t}$, *it holds that*

$$|\mathcal{V}(\pi) - \mathcal{V}(\pi, \mathbf{b}_R, \mathbf{b}_D)| \leq \sum_{t=1}^{T} C_t^{\pi} \sqrt{\mathrm{vol}(\mathcal{R} \times \mathcal{O})} \sqrt{\mathcal{L}_R^{[t]}(b_{R,t})} + \sum_{t=1}^{T-1} C_t^{\pi} (T-t) \sqrt{\mathrm{vol}(\mathcal{O} \times \mathcal{O})} \sqrt{\mathcal{L}_D^{[t]}(b_{D,t})}, \tag{53}$$

where C_t^{π} *is defined in Definition A.3, and* $\mathcal{L}_R^{[t]}$, $\mathcal{L}_D^{[t]}$ *denotes the risk functionals defined in Definition A.2.*

Proof. The proof is provided in Appendix D.3. □

The next lemma essentially shows that the true bridge functions are contained in the constructed confidence sets with high probability.

Lemma C.2. *Under Assumptions 3.1, 3.2, 3.3, 4.1, D.16, with probability at least* $1-\delta$, *for some* $c > 0$, *by setting* $\lambda_1 = N_1^{-\frac{1}{c_1+1}}$, $\lambda_2 = N_2^{-\frac{\gamma}{\gamma c_2+1}}$, $N_1 = N_2^{\frac{c_1+1}{c_1-1}\frac{\gamma(c_2+1)}{\gamma c_2+1}}$, *and setting*

$$\alpha = c\log(T/\delta) M_R N_2^{-\frac{\gamma}{2\gamma+2}}, \beta = c\log(T/\delta) M_D N_2^{-\frac{\gamma}{2\gamma+2}},$$

it holds that

$$(b_R^{[1]}, ..., b_R^{[T]}) \in \mathrm{conf}_R(\alpha),\ (b_D^{[1]}, ..., b_D^{[T-1]}) \in \mathrm{conf}_D(\beta). \tag{54}$$

Here γ, *are defined in Assumption 4.1. Constants* c_1, c_2 *denote a measure of smoothness and are defined in Assumption D.10, D.16 in Appendix D.7.*

Proof. See Appendix D.5 for a proof. □

Next, the following lemma requires a uniform upper bound within the confidence regions.

Lemma C.3. *Under Assumptions 3.1, 3.2, 3.3, 4.1, D.16, with probability at least* $1-\delta$, *for some* $c > 0$, *by setting* $\lambda_1 = N_1^{-\frac{1}{c_1+1}}$, $\lambda_2 = N_2^{-\frac{\gamma}{\gamma c_2+1}}$, $N_1 = N_2^{\frac{c_1+1}{c_1-1}\frac{\gamma(c_2+1)}{\gamma c_2+1}}$, *and setting*

$$\alpha = c\log(T/\delta) M_R N_2^{-\frac{\gamma}{2\gamma+2}}, \beta = c\log(T/\delta) M_D N_2^{-\frac{\gamma}{2\gamma+2}},$$

it holds that

$$\sup_{\mathbf{b}_R \in \mathrm{conf}_R(\alpha)} \max_{t=1:T} \sqrt{\mathcal{L}_R^{[t]}(b_{R,t})} \leq \sqrt{\alpha}, \tag{55}$$

$$\sup_{\mathbf{b}_D \in \mathrm{conf}_D(\beta)} \max_{t=1:T-1} \sqrt{\mathcal{L}_D^{[t]}(b_{D,t})} \leq \sqrt{\beta}. \tag{56}$$

Here γ, M_R, M_D *are defined in Assumption 4.1. Constants* c_1, c_2 *denote a measure of smoothness and are defined in Assumption D.10, D.16 in Appendix D.7.*

Proof. See Appendix D.5 for a proof. □

Then, by combining Lemma C.1, Lemma C.2 and Lemma C.3, we obtain the following upper bounds

$$\mathcal{V}(\pi^\star) - \mathcal{V}(\widehat{\pi})$$

$$
\begin{aligned}
&=\mathcal{V}(\pi^\star) - \min_{(\mathbf{b}_R,\mathbf{b}_D)\in \text{conf}_R(\alpha)\times\text{conf}_D(\beta)} \mathcal{V}(\pi^\star,\mathbf{b}_R,\mathbf{b}_D) + \min_{(\mathbf{b}_R,\mathbf{b}_D)\in \text{conf}_R(\alpha)\times\text{conf}_D(\beta)} \mathcal{V}(\pi^\star,\mathbf{b}_R,\mathbf{b}_D) - \mathcal{V}(\widehat{\pi}) \\
&\leq\mathcal{V}(\pi^\star) - \min_{(\mathbf{b}_R,\mathbf{b}_D)\in \text{conf}_R(\alpha)\times\text{conf}_D(\beta)} \mathcal{V}(\pi^\star,\mathbf{b}_R,\mathbf{b}_D) + \min_{(\mathbf{b}_R,\mathbf{b}_D)\in \text{conf}_R(\alpha)\times\text{conf}_D(\beta)} \mathcal{V}(\widehat{\pi},\mathbf{b}_R,\mathbf{b}_D) - \mathcal{V}(\widehat{\pi}) \\
&\leq\mathcal{V}(\pi^\star) - \min_{(\mathbf{b}_R,\mathbf{b}_D)\in \text{conf}_R(\alpha)\times\text{conf}_D(\beta)} \mathcal{V}(\pi^\star,\mathbf{b}_R,\mathbf{b}_D) + 0 \\
&\leq \sup_{(\mathbf{b}_R,\mathbf{b}_D)\in \text{conf}_R(\alpha)\times\text{conf}_D(\beta)} \{\sum_{t=1}^{T} C_t^{\pi^*}\sqrt{\text{vol}(\mathcal{R}\times\mathcal{O})}\sqrt{\mathcal{L}_R^{[t]}(b_{R,t})} + \sum_{t=1}^{T-1} C_t^{\pi^*}(T-t)\sqrt{\text{vol}(\mathcal{O}\times\mathcal{O})}\sqrt{\mathcal{L}_D^{[t]}(b_{D,t})}\} \\
&\lesssim \sum_{t=1}^{T} C_t^{\pi^*}\sqrt{\log(T/\delta)M_R}N_2^{-\frac{\gamma}{4\gamma+4}} + \sum_{t=1}^{T-1} C_t^{\pi^*}(T-t)\sqrt{\log(T/\delta)M_D}N_2^{-\frac{\gamma}{4\gamma+4}} \\
&=(\sum_{t=1}^{T}(\sqrt{M_R}+(T-t)\sqrt{M_D})C_t^{\pi^*})\sqrt{\log(T/\delta)}N_2^{-\frac{\gamma}{4\gamma+4}}
\end{aligned} \tag{57}
$$

The first inequality is by the definition of $\widehat{\pi}$. The second inequality is from Lemma C.2. The third inequality is from Lemma C.1. The final inequality is from Lemma C.3.

The proof of Theorem 4.2 is completed.

D. Proof for Lemmas

D.1. Proof of Lemma A.1

Proof.

$$
\begin{aligned}
&\mathbb{E}[b(W,y) \mid X=x] \\
=&\int_{\mathcal{W}} b(w,y)p(w\mid x)dw \\
=&\int_{\mathcal{W}} \langle b, \phi(w)\otimes\phi(y)\rangle_{\mathcal{H}_\mathcal{W}\otimes\mathcal{H}_\mathcal{Y}} p(w\mid x)dw \\
=&\langle b, \int_{\mathcal{W}} \phi(w)p(w\mid x)dw \otimes \phi(y)\rangle_{\mathcal{H}_\mathcal{W}\otimes\mathcal{H}_\mathcal{Y}} \\
=&\langle b, \mu_{W|x}\otimes\phi(y)\rangle_{\mathcal{H}_\mathcal{W}\otimes\mathcal{H}_\mathcal{Y}}
\end{aligned} \tag{58}
$$

where we have used the Bochner integrability of the feature map $\phi(w)$ to take the expectation inside the dot product. See Definition A.5.20 for more details (Steinwart & Christmann, 2008). □

D.2. Proof of Lemma B.1

Proof of Lemma B.1. Let $\bar{H}_t := (S_1, O_1, A_1, ..., S_t, O_t, A_t)$ and $H_t := (O_1, A_1, ..., O_t, A_t)$. We have

$$
\begin{aligned}
&p^\pi(r_{t+1}) \\
=&\int p^\pi(r_{t+1}\mid \bar{h}_{t+1})p^\pi(\bar{h}_{t+1}) \\
=&\int p(r_{t+1}\mid s_{t+1},a_{t+1})p^\pi(\bar{h}_{t+1}) \\
=&\sum_{a_{t+1}}\int_{o_{t+1},s_{t+1},\bar{h}_t} p(r_{t+1}\mid s_{t+1},a_{t+1})\pi(a_{t+1}\mid o_{t+1},h_t)p(o_{t+1}\mid s_{t+1})p(s_{t+1}\mid s_t,a_t)p^\pi(\bar{h}_t) \\
=&\cdots \\
=&\sum_{a_{t+1},a_t,\ldots,a_1}\int_{o_{t+1},s_{t+1},\ldots,o_1,s_1} p(r_{t+1}\mid s_{t+1},a_{t+1})\prod_{i=1}^{t+1}\{\pi(a_i\mid o_i,h_{i-1})p(o_i\mid s_i)p(s_i\mid s_{i-1},a_{i-1})\}
\end{aligned}
$$

$$= \sum_{a_{t+1},a_t,\ldots,a_1} \int_{o_{t+1},\ldots,o_1} \prod_{i=1}^{t+1} \pi(a_i \mid o_i, h_{i-1}) \int_{s_{t+1},\ldots,s_1} p(r_{t+1}, o_{t+1} \mid s_{t+1}, a_{t+1}) \prod_{i=1}^{t} p(s_{i+1}, o_i \mid s_i, a_i) p(s_1) \tag{59}$$

□

D.3. Proof of Lemma C.1

We need to upper bound $|\mathcal{V}(\pi) - \mathcal{V}(\pi, \mathbf{b}_R, \mathbf{b}_D)|$, for any $\pi \in \Pi$, $\mathbf{b}_R \in \otimes_{t=1}^{T} \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}$, $\mathbf{b}_D \in \otimes_{t=1}^{T-1} \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Z}_t}$.

According to Theorem 3.5, we have

$$f_t(r_t, h_t) = \int_{\tilde{o}_t,\tilde{o}_{t-1},\ldots,\tilde{o}_1} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1), \tag{60}$$

and that $p^\pi(r_t)$ can be identified by

$$p^\pi(r_t) = \sum_{a_t,a_{t-1},\ldots,a_1} \int_{o_t,o_{t-1},\ldots,o_1} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, a_{j-1}, o_{j-1}, \ldots) f_t(r_t, h_t). \tag{61}$$

In this section, we also define two other notions:

$$\widetilde{f}_t(r_t, h_t) = \int_{\tilde{o}_t,\tilde{o}_{t-1},\ldots,\tilde{o}_1} b_{R,t}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1), \tag{62}$$

and

$$\widehat{f}_t(r_t, h_t) = \int_{\tilde{o}_t,\tilde{o}_{t-1},\ldots,\tilde{o}_1} b_{R,t}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{t-1} b_{D,j}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1). \tag{63}$$

In $\widetilde{f}_t$, the true reward-emission bridge function $b_R^{[t]}$ is replaced by a generic $b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Z}_t}$. In $\widehat{f}_t$, all the true bridge functions are replaced by generic elements $b_{R,t}$ and $\{b_{D,j}\}_{j=1}^{t-1}$.

Similarly, we define

$$\widetilde{p}_t^\pi(r_t) = \sum_{a_t,a_{t-1},\ldots,a_1} \int_{o_t,o_{t-1},\ldots,o_1} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, a_{j-1}, o_{j-1}, \ldots) \widetilde{f}_t(r_t, h_t), \tag{64}$$

$$\widehat{p}_t^\pi(r_t) = \sum_{a_t,a_{t-1},\ldots,a_1} \int_{o_t,o_{t-1},\ldots,o_1} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, a_{j-1}, o_{j-1}, \ldots) \widehat{f}_t(r_t, h_t). \tag{65}$$

Next, we observe that

$$\begin{aligned}
&\mathcal{V}(\pi) - \mathcal{V}(\pi, \mathbf{b}_R, \mathbf{b}_D) \\
=&\sum_{t=1}^{T} \int_{r_t} r_t p_t^\pi(r_t) - \sum_{t=1}^{T} \int_{r_t} r_t \widehat{p}_t^\pi(r_t) \\
=&\sum_{t=1}^{T} \int_{r_t} r_t p_t^\pi(r_t) - \sum_{t=1}^{T} \int_{r_t} r_t \widetilde{p}_t^\pi(r_t) + \sum_{t=1}^{T} \int_{r_t} r_t \widetilde{p}_t^\pi(r_t) - \sum_{t=1}^{T} \int_{r_t} r_t \widehat{p}_t^\pi(r_t) \\
=&\mathcal{V}(\pi) - \widetilde{\mathcal{V}}(\pi, \mathbf{b}_R) + \widetilde{\mathcal{V}}(\pi, \mathbf{b}_R) - \mathcal{V}(\pi, \mathbf{b}_R, \mathbf{b}_D)
\end{aligned} \tag{66}$$

where we denote $\sum_{t=1}^{T} \int_{r_t} r_t \widetilde{p}_t^\pi(r_t)$ as $\widetilde{\mathcal{V}}(\pi, \mathbf{b}_R)$.

In the rest of the proof, we provide upper bounds on $\mathcal{V}(\pi) - \widetilde{\mathcal{V}}(\pi, \mathbf{b}_R)$ and $\widetilde{\mathcal{V}}(\pi, \mathbf{b}_R) - \mathcal{V}(\pi, \mathbf{b}_R, \mathbf{b}_D)$ separately.

Bounding $\mathcal{V}(\pi) - \widetilde{\mathcal{V}}(\pi, \mathbf{b}_R)$.

By definition, we have

$$\begin{aligned}
&\mathcal{V}(\pi) - \widetilde{\mathcal{V}}(\pi, \mathbf{b}_R) \\
=&\sum_{t=1}^{T} \int_{r_t} r_t p_t^{\pi}(r_t) - \sum_{t=1}^{T} \int_{r_t} r_t \widetilde{p}_t^{\pi}(r_t) \\
=&\sum_{t=1}^{T} \int_{r_t} r_t (p_t^{\pi}(r_t) - \widetilde{p}_t^{\pi}(r_t))
\end{aligned} \tag{67}$$

where

$$\begin{aligned}
&p_t^{\pi}(r_t) - \widetilde{p}_t^{\pi}(r_t) \\
=&\sum_{a_t, a_{t-1}, \ldots, a_1} \int_{o_t, o_{t-1}, \ldots, o_1} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, a_{j-1}, o_{j-1}, \ldots)(f_t(r_t, h_t) - \widetilde{f}_t(r_t, h_t))
\end{aligned} \tag{68}$$

with

$$\begin{aligned}
&f_t(r_t, h_t) - \widetilde{f}_t(r_t, h_t) \\
=&\int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_1} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) - \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_1} b_{R,t}(a_t, \tilde{o}_t, r_t, o_t) \\
&\prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
=&\int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_1} (b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) - b_{R,t}(a_t, \tilde{o}_t, r_t, o_t)) \prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1).
\end{aligned} \tag{69}$$

Then, we apply the same strategy from the proof of Theorem 3.5. In particular, recall that we have shown

$$\begin{aligned}
&f_t(r_t, h_t) \\
=&\int_{s_t, s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) p(\tilde{o}_t \mid s_t) \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
=&\int_{\tilde{o}_t, \ldots \tilde{o}_1} b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1).
\end{aligned} \tag{70}$$

By the same arguments, it is also straightforward to have

$$\begin{aligned}
&\int_{\tilde{o}_t, \ldots \tilde{o}_1} b_{R,t}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
=&\int_{s_t, s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} b_{R,t}(a_t, \tilde{o}_t, r_t, o_t) p(\tilde{o}_t \mid s_t) \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1).
\end{aligned} \tag{71}$$

We use $\Delta(b_R^{[t]}, b_{R,t})$ to denote $b_R^{[t]} - b_{R,t}$. Then, we have

$$\begin{aligned}
&f_t(r_t, h_t) - \widetilde{f}_t(r_t, h_t) \\
&= \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_1} (b_R^{[t]}(a_t, \tilde{o}_t, r_t, o_t) - b_{R,t}(a_t, \tilde{o}_t, r_t, o_t)) \prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
&= \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_1} \Delta(b_R^{[t]}, b_{R,t})(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{t-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
&= \int_{s_t, s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} \Delta(b_R^{[t]}, b_{R,t})(a_t, \tilde{o}_t, r_t, o_t) p(\tilde{o}_t \mid s_t) \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1)
\end{aligned} \tag{72}$$

where the last equality comes from (70)(71).

Therefore, we have

$$\begin{aligned}
&p_t^\pi(r_t) - \widetilde{p}_t^\pi(r_t) \\
&= \sum_{a_t, a_{t-1}, \ldots, a_1} \int_{o_t, o_{t-1}, \ldots, o_1} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, a_{j-1}, o_{j-1}, \ldots)(f_t(r_t, h_t) - \widetilde{f}_t(r_t, h_t)) \\
&= \sum_{a_t, a_{t-1}, \ldots, a_1} \int_{o_t, o_{t-1}, \ldots, o_1} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, h_{j-1}) \int_{s_t, s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} \Delta(b_R^{[t]}, b_{R,t})(a_t, \tilde{o}_t, r_t, o_t) p(\tilde{o}_t \mid s_t) \\
&\quad \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1)
\end{aligned} \tag{73}$$

We need the following lemma to proceed.

Lemma D.1. *For any measurable function $g(a_t, \tilde{o}_t, h_{t-1}, o_t)$: $\mathcal{A} \times \mathcal{O} \times \mathcal{H}_{t-1} \times \mathcal{O} \to \mathbb{R}$, it holds that*

$$\begin{aligned}
&\sum_{a_t, a_{t-1}, \ldots, a_1} \int_{o_t, o_{t-1}, \ldots, o_1} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, h_{j-1}) \int_{s_t, s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} g(a_t, \tilde{o}_t, h_{t-1}, o_t) p(\tilde{o}_t \mid s_t) \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
&= \int_{o_t} \mathbb{E}\left[\frac{p^\pi(S_t, H_{t-1}) \pi_t(A_t \mid o_t, H_{t-1})}{p^{\pi^b}(S_t, H_{t-1}) \pi_t^b(A_t \mid S_t)} g(A_t, O_t, H_{t-1}, o_t) \right].
\end{aligned} \tag{74}$$

According to Lemma D.1, we can further express $p_t^\pi(r_t) - \widetilde{p}_t^\pi(r_t)$ as

$$\begin{aligned}
&p_t^\pi(r_t) - \widetilde{p}_t^\pi(r_t) \\
&= \sum_{a_t, a_{t-1}, \ldots, a_1} \int_{o_t, o_{t-1}, \ldots, o_1} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, h_{j-1}) \int_{s_t, s_{t-1}, \ldots, s_1} \int_{\tilde{o}_t} \Delta(b_R^{[t]}, b_{R,t})(a_t, \tilde{o}_t, r_t, o_t) p(\tilde{o}_t \mid s_t) \\
&\quad \times \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
&= \int_{o_t} \mathbb{E}\left[\frac{p^\pi(S_t, H_{t-1}) \pi_t(A_t \mid o_t, H_{t-1})}{p^{\pi^b}(S_t, H_{t-1}) \pi_t^b(A_t \mid S_t)} \Delta(b_R^{[t]}, b_{R,t})(A_t, O_t, r_t, o_t) \right].
\end{aligned} \tag{75}$$

Note that in the application of Lemma D.1 in the above derivation, we let $g(a_t, \tilde{o}_t, h_{t-1}, o_t)$ be the function $\Delta(b_R^{[t]}, b_{R,t})(a_t, \tilde{o}_t, r_t, o_t)$, where the input of h_{t-1} is empty, and r_t is a fixed number (i.e. not a variable).

Consequently, we have

$$\begin{aligned}
&\mathcal{V}(\pi)-\widetilde{\mathcal{V}}(\pi,\mathbf{b}_R)\\
=&\sum_{t=1}^{T}\int_{r_t} r_t(p_t^{\pi}(r_t)-\widetilde{p}_t^{\pi}(r_t))\\
=&\sum_{t=1}^{T}\int_{r_t,o_t} r_t\mathbb{E}\left[\frac{p^{\pi}\left(S_t,H_{t-1}\right)\pi_t\left(A_t\mid o_t,H_{t-1}\right)}{p^{\pi^b}\left(S_t,H_{t-1}\right)\pi_t^b\left(A_t\mid S_t\right)}\Delta(b_R^{[t]},b_{R,t})\left(A_t,O_t,r_t,o_t\right)\right] \text{ by (75).}
\end{aligned} \tag{76}$$

Next, we upper bound $\mathcal{V}(\pi)-\widetilde{\mathcal{V}}(\pi,\mathbf{b}_R)$ based on (76). For each t, we have

$$\begin{aligned}
&\int_{r_t,o_t} r_t\mathbb{E}\left[\frac{p^{\pi}\left(S_t,H_{t-1}\right)\pi_t\left(A_t\mid o_t,H_{t-1}\right)}{p^{\pi^b}\left(S_t,H_{t-1}\right)\pi_t^b\left(A_t\mid S_t\right)}\Delta(b_R^{[t]},b_{R,t})\left(A_t,O_t,r_t,o_t\right)\right]\\
=&\int_{r_t,o_t} r_t\mathbb{E}[\mathbb{E}[w_t^{\pi}(A_t,O_0,H_{t-1},o_t)\mid S_t,A_t,H_{t-1}]\Delta(b_R^{[t]},b_{R,t})(A_t,O_t,r_t,o_t)] \text{ by Assumption 3.2}\\
=&\int_{r_t,o_t} r_t\mathbb{E}[\mathbb{E}[w_t^{\pi}(A_t,O_0,H_{t-1},o_t)\mid S_t,A_t,H_{t-1},O_t]\Delta(b_R^{[t]},b_{R,t})(A_t,O_t,r_t,o_t)] \text{ by Assumption 3.1}\\
=&\int_{r_t,o_t} r_t\mathbb{E}[\mathbb{E}[w_t^{\pi}(A_t,O_0,H_{t-1},o_t)\Delta(b_R^{[t]},b_{R,t})(A_t,O_t,r_t,o_t)\mid S_t,A_t,H_{t-1},O_t]] \text{ by measurability}\\
=&\int_{r_t,o_t} r_t\mathbb{E}[w_t^{\pi}(A_t,O_0,H_{t-1},o_t)\Delta(b_R^{[t]},b_{R,t})(A_t,O_t,r_t,o_t)]\\
=&\int_{r_t,o_t} r_t\mathbb{E}[\mathbb{E}[w_t^{\pi}(A_t,O_0,H_{t-1},o_t)\Delta(b_R^{[t]},b_{R,t})(A_t,O_t,r_t,o_t)\mid A_t,H_{t-1},O_0]]\\
=&\int_{r_t,o_t} r_t\mathbb{E}[w_t^{\pi}(A_t,O_0,H_{t-1},o_t)\mathbb{E}[\Delta(b_R^{[t]},b_{R,t})(A_t,O_t,r_t,o_t)\mid A_t,H_{t-1},O_0]]\\
\leq&\int_{r_t,o_t}\sqrt{\mathbb{E}[w_t^{\pi}(A_t,H_{t-1},O_0,o_t)^2]}\sqrt{\mathbb{E}[\{\mathbb{E}[\Delta(b_R^{[t]},b_{R,t})(A_t,O_t,r_t,o_t)\mid A_t,H_{t-1},O_0]\}^2]}\\
&\text{(Cauchy-Schwartz inequality and } |r_t|\leq 1)\\
\leq&\sqrt{\int_{o_t}\mathbb{E}[w_t^{\pi}(A_t,H_{t-1},O_0,o_t)^2]}\sqrt{\int_{r_t,o_t}\mathbb{E}[\{\mathbb{E}[\Delta(b_R^{[t]},b_{R,t})(A_t,O_t,r_t,o_t)\mid A_t,H_{t-1},O_0]\}^2]}\\
&\text{(Cauchy-Schwartz inequality)}\\
=&C_t^{\pi}\sqrt{\int_{r_t,o_t}\mathbb{E}[\{\mathbb{E}[\Delta(b_R^{[t]},b_{R,t})(A_t,O_t,r_t,o_t)\mid A_t,H_{t-1},O_0]\}^2]} \text{ by definition of concentrability coefficient}\\
=&C_t^{\pi}\sqrt{\int_{r_t,o_t}\|p(r_t,o_t\mid A_t,H_{t-1},O_0)-\mathbb{E}[b_{R,t}(A_t,O_t,r_t,o_t)\mid A_t,H_{t-1},O_0]\|^2_{\mathcal{L}^2(\mathbb{P}^{\pi^b})}} \text{ by Assumption 3.2}\\
=&C_t^{\pi}\sqrt{\mathrm{vol}(\mathcal{R}\times\mathcal{O})}\sqrt{\mathcal{L}_R^{[t]}(b_{R,t})} \text{ by Definition A.2.}
\end{aligned} \tag{77}$$

Consequently, we can upper-bound $\mathcal{V}(\pi)-\widetilde{\mathcal{V}}(\pi,\mathbf{b}_R)$ by

$$\mathcal{V}(\pi)-\widetilde{\mathcal{V}}(\pi,\mathbf{b}_R)\leq\sum_{t=1}^{T}C_t^{\pi}\sqrt{\mathrm{vol}(\mathcal{R}\times\mathcal{O})}\sqrt{\mathcal{L}_R^{[t]}(b_{R,t})} \tag{78}$$

through combining (76)(77). In addition, by symmetry, we also have

$$\widetilde{\mathcal{V}}(\pi,\mathbf{b}_R)-\mathcal{V}(\pi)\leq\sum_{t=1}^{T}C_t^{\pi}\sqrt{\mathrm{vol}(\mathcal{R}\times\mathcal{O})}\sqrt{\mathcal{L}_R^{[t]}(b_{R,t})}. \tag{79}$$

Bounding $\widetilde{\mathcal{V}}(\pi, \mathbf{b}_R) - \mathcal{V}(\pi, \mathbf{b}_R, \mathbf{b}_D)$.

Firstly, we introduce a notion as follows

$$f_{t,i}(r_t, h_t) = \int_{\tilde{o}_t, \tilde{o}_{t-1}, ..., \tilde{o}_1} b_{R,t}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{\min\{i,t\}-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) \prod_{j=\min\{i,t\}}^{t-1} b_{D,j}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1). \tag{80}$$

where $\prod_{j=p}^{q} u_j := 1$ if $p > q$. Notice that $f_{t,T} = \widetilde{f}_t$ and $f_{t,1} = \widehat{f}_t$.

Similarly, we define

$$p_{t,i}^{\pi}(r_t) = \sum_{a_t, a_{t-1}, ..., a_1} \int_{o_t, o_{t-1}, ..., o_1} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, a_{j-1}, o_{j-1}, ...) f_{t,i}(r_t, h_t), \tag{81}$$

and notice that $p_{t,T}^{\pi} = \widetilde{p}_t^{\pi}$, $p_{t,1}^{\pi} = \widehat{p}_t^{\pi}$.

We then define

$$G_i = \sum_{t=1}^{T} \int_{r_t} r_t p_{t,i}^{\pi}(r_t) \tag{82}$$

and notice that

$$G_1 = \sum_{t=1}^{T} \int_{r_t} r_t p_{t,1}^{\pi}(r_t) = \sum_{t=1}^{T} \int_{r_t} r_t \widehat{p}_t^{\pi}(r_t) = \mathcal{V}(\pi, \mathbf{b}_R, \mathbf{b}_D) \tag{83}$$

and

$$G_T = \sum_{t=1}^{T} \int_{r_t} r_t p_{t,T}^{\pi}(r_t) = \sum_{t=1}^{T} \int_{r_t} r_t \widetilde{p}_t^{\pi}(r_t) = \widetilde{\mathcal{V}}(\pi, \mathbf{b}_R). \tag{84}$$

Therefore, we have

$$\widetilde{\mathcal{V}}(\pi, \mathbf{b}_R) - \mathcal{V}(\pi, \mathbf{b}_R, \mathbf{b}_D) = G_T - G_1 = \sum_{i=1}^{T-1} (G_{i+1} - G_i) \tag{85}$$

Next, we focus on the term $G_{i+1} - G_i$ for each $i = 1, ..., T-1$. We first write

$$G_{i+1} - G_i = \sum_{t=1}^{T} \int_{r_t} r_t p_{t,i+1}^{\pi}(r_t) - \sum_{t=1}^{T} \int_{r_t} r_t p_{t,i}^{\pi}(r_t). \tag{86}$$

We notice that if $t \leq i$, then $p_{t,i+1}^{\pi}(r_t) = p_{t,i}^{\pi}(r_t)$ for each r_t. It is because both $f_{t,i+1}(r_t, h_t)$ and $f_{t,i}(r_t, h_t)$ are equal to $f_t(r_t, h_t)$. Thus, $G_{i+1} - G_i$ can be simplified as

$$\begin{aligned}
& G_{i+1} - G_i \\
&= \sum_{t=i+1}^{T} \int_{r_t} r_t p_{t,i+1}^{\pi}(r_t) - \sum_{t=i+1}^{T} \int_{r_t} r_t p_{t,i}^{\pi}(r_t) \\
&= \sum_{t=i+1}^{T} \int_{r_t} r_t (p_{t,i+1}^{\pi}(r_t) - p_{t,i}^{\pi}(r_t)) \\
&= \sum_{t=i+1}^{T} \int_{r_t} r_t \int_{h_t} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, h_{j-1}) (f_{t,i+1}(r_t, h_t) - f_{t,i}(r_t, h_t))
\end{aligned} \tag{87}$$

where

$$
\begin{aligned}
& f_{t,i+1}(r_t, h_t) - f_{t,i}(r_t, h_t) \\
= & \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_1} b_{R,t}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{\min\{i+1,t\}-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) \prod_{j=\min\{i+1,t\}}^{t-1} b_{D,j}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
- & \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_1} b_{R,t}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{\min\{i,t\}-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) \prod_{j=\min\{i,t\}}^{t-1} b_{D,j}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
= & \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_1} b_{R,t}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{i} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) \prod_{j=i+1}^{t-1} b_{D,j}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
- & \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_1} b_{R,t}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{i-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) \prod_{j=i}^{t-1} b_{D,j}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
= & \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_1} b_{R,t}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{i-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) b_D^{[i]}(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i) \prod_{j=i+1}^{t-1} b_{D,j}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
- & \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_1} b_{R,t}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=1}^{i-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) b_{D,i}(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i) \prod_{j=i+1}^{t-1} b_{D,j}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1)
\end{aligned} \tag{88}
$$

Next, we introduce the following definition.

Definition D.2 (Value function)**.** For each $t = 1, ..., T$ and each $i \leq t-1$, we define

$$
\begin{aligned}
& u_t(\tilde{o}_{i+1}, h_i) \\
= & \int_{r_t} r_t \int_{a_{i+1}, o_{i+1}, \ldots, o_t, a_t} \prod_{j=i+1}^{t} \pi_j(a_j \mid o_j, h_{j-1}) \int_{\tilde{o}_t, \tilde{o}_{t-1}, \ldots, \tilde{o}_{i+2}} b_{R,t}(a_t, \tilde{o}_t, r_t, o_t) \prod_{j=i+1}^{t-1} b_{D,j}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j),
\end{aligned} \tag{89}
$$

and

$$
U_{i+1}(\tilde{o}_{i+1}, h_i) := \sum_{t=i+1}^{T} u_t(\tilde{o}_{i+1}, h_i). \tag{90}
$$

Intuitively, $U_{i+1}(\tilde{o}_{i+1}, h_i)$ in Definition D.2 can be understood as a kind of value function in some sense, which plays a similar role with the value function in the standard MDP settings at the stage $i+1$. In the standard MDP settings, the value function at the stage $i+1$ is upper bounded by $T-i$ when the reward function satisfies $r_t \in [-1, 1]$. Therefore, in this work, it is natural to assume that $|U_{i+1}(\tilde{o}_{i+1}, h_i)| \leq T - i$ for every $i = 0, ..., T-1$, which is summarized in Assumption D.3.

Assumption D.3. For each $i = 1, ..., T$, it holds for all $(\tilde{o}_i, h_{i-1}) \in \mathcal{O} \times \mathcal{H}_{i-1}$ that

$$
|U_i(\tilde{o}_i, h_{i-1})| \leq T - i + 1. \tag{91}
$$

We also use the following notation to denote the difference between the true dynamic-emission bridge function and a generic element:

$$
\Delta(b_D^{[i]}, b_{D,i})(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i) := b_D^{[i]}(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i) - b_{D,i}(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i). \tag{92}
$$

By incorporating Definition D.2 and the notation $\Delta(b_D^{[i]}, b_{D,i})$, we then have

$$
\begin{aligned}
&G_{i+1} - G_i \\
&= \sum_{t=i+1}^{T} \int_{r_t} r_t \int_{h_t} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, h_{j-1})(f_{t,i+1}(r_t, h_t) - f_{t,i}(r_t, h_t)) \\
&= \sum_{t=i+1}^{T} \int_{h_i} \int_{\tilde{o}_{i+1},\ldots,\tilde{o}_1} u_t(\tilde{o}_{i+1}, h_i) \prod_{j=1}^{i} \pi_j(a_j \mid o_j, h_{j-1}) b_D^{[i]}(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i) \prod_{j=1}^{i-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
&- \sum_{t=i+1}^{T} \int_{h_i} \int_{\tilde{o}_{i+1},\ldots,\tilde{o}_1} u_t(\tilde{o}_{i+1}, h_i) \prod_{j=1}^{i} \pi_j(a_j \mid o_j, h_{j-1}) b_{D,i}(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i) \prod_{j=1}^{i-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
&= \int_{h_i} \int_{\tilde{o}_{i+1},\ldots,\tilde{o}_1} (\sum_{t=i+1}^{T} u_t(\tilde{o}_{i+1}, h_i)) \prod_{j=1}^{i} \pi_j(a_j \mid o_j, h_{j-1}) b_D^{[i]}(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i) \prod_{j=1}^{i-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
&- \int_{h_i} \int_{\tilde{o}_{i+1},\ldots,\tilde{o}_1} (\sum_{t=i+1}^{T} u_t(\tilde{o}_{i+1}, h_i)) \prod_{j=1}^{i} \pi_j(a_j \mid o_j, h_{j-1}) b_{D,i}(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i) \prod_{j=1}^{i-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
&= \int_{h_i} \prod_{j=1}^{i} \pi_j(a_j \mid o_j, h_{j-1}) \int_{\tilde{o}_{i+1},\ldots,\tilde{o}_1} U_{i+1}(\tilde{o}_{i+1}, h_i) \Delta(b_D^{[i]}, b_{D,i})(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i) \prod_{j=1}^{i-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1)
\end{aligned}
\tag{93}
$$

We first focus on the term $\int_{\tilde{o}_{i+1},\ldots,\tilde{o}_1} U_{i+1}(\tilde{o}_{i+1}, h_i) \Delta(b_D^{[i]}, b_{D,i})(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i) \prod_{j=1}^{i-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1)$ in the last equality of (93). By sequentially applying the definition of $b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j)$ and the same arguments used in the proof of Theorem 3.5, we have

$$
\begin{aligned}
&\int_{\tilde{o}_{i+1},\ldots,\tilde{o}_1} U_{i+1}(\tilde{o}_{i+1}, h_i) \Delta(b_D^{[i]}, b_{D,i})(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i) \prod_{j=1}^{i-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
&= \int_{s_i, s_{i-1},\ldots,s_1} \int_{\tilde{o}_{i+1}, \tilde{o}_i} U_{i+1}(\tilde{o}_{i+1}, h_i) \Delta(b_D^{[i]}, b_{D,i})(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i) p(\tilde{o}_i \mid s_i) \prod_{j=1}^{i-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1).
\end{aligned}
\tag{94}
$$

Then, by applying Lemma D.1 again for the function $\int_{\tilde{o}_{i+1}} U_{i+1}(\tilde{o}_{i+1}, h_i) \Delta(b_D^{[i]}, b_{D,i})(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i)$, we have

$$
\begin{aligned}
&G_{i+1} - G_i \\
&= \int_{h_i} \prod_{j=1}^{i} \pi_j(a_j \mid o_j, h_{j-1}) \int_{\tilde{o}_{i+1},\ldots,\tilde{o}_1} U_{i+1}(\tilde{o}_{i+1}, h_i) \Delta(b_D^{[i]}, b_{D,i})(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i) \prod_{j=1}^{i-1} b_D^{[j]}(a_j, \tilde{o}_j, \tilde{o}_{j+1}, o_j) p(\tilde{o}_1) \\
&= \int_{h_i} \prod_{j=1}^{i} \pi_j(a_j \mid o_j, h_{j-1}) \int_{s_i, s_{i-1},\ldots,s_1} \int_{\tilde{o}_{i+1}, \tilde{o}_i} U_{i+1}(\tilde{o}_{i+1}, h_i) \Delta(b_D^{[i]}, b_{D,i})(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i) p(\tilde{o}_i \mid s_i) \\
&\prod_{j=1}^{i-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1) \\
&= \int_{o_i} \mathbb{E}\left[\frac{p^{\pi}(S_i, H_{i-1}) \pi_i(A_i \mid o_i, H_{i-1})}{p^{\pi^b}(S_i, H_{i-1}) \pi_i^b(A_i \mid S_i)} \int_{\tilde{o}_{i+1}} U_{i+1}(\tilde{o}_{i+1}, A_i, o_i, H_{i-1}) \Delta(b_D^{[i]}, b_{D,i})(A_i, O_i, \tilde{o}_{i+1}, o_i) \right].
\end{aligned}
\tag{95}
$$

In the above derivation, the function $g(a_i, \tilde{o}_i, h_{i-1}, o_i)$ is defined as $\int_{\tilde{o}_{i+1}} U_{i+1}(\tilde{o}_{i+1}, h_i) \Delta(b_D^{[i]}, b_{D,i})(a_i, \tilde{o}_i, \tilde{o}_{i+1}, o_i)$ when we apply Lemma D.1.

Next, we provide an upper bound on the last equality of (95). For clarity, we use the notation $\kappa(A_i, O_i, H_{i-1}, \tilde{o}_{i+1}, o_i)$ to

denote $U_{i+1}(\tilde{o}_{i+1}, A_i, o_i, H_{i-1})\Delta(b_D^{[i]}, b_{D,i})(A_i, O_i, \tilde{o}_{i+1}, o_i)$. Then, we have

$$
\begin{aligned}
&G_{i+1} - G_i \\
=&\int_{o_i,\tilde{o}_{i+1}} \mathbb{E}\left[\frac{p^{\pi}\left(S_i, H_{i-1}\right)\pi_i\left(A_i \mid o_i, H_{i-1}\right)}{p^{\pi^b}\left(S_i, H_{i-1}\right)\pi_i^b\left(A_i \mid S_i\right)} U_{i+1}(\tilde{o}_{i+1}, A_i, o_i, H_{i-1})\Delta(b_D^{[i]}, b_{D,i})(A_i, O_i, \tilde{o}_{i+1}, o_i)\right] \\
=&\int_{o_i,\tilde{o}_{i+1}} \mathbb{E}\left[\frac{p^{\pi}\left(S_i, H_{i-1}\right)\pi_i\left(A_i \mid o_i, H_{i-1}\right)}{p^{\pi^b}\left(S_i, H_{i-1}\right)\pi_i^b\left(A_i \mid S_i\right)} \kappa(A_i, O_i, H_{i-1}, \tilde{o}_{i+1}, o_i)\right] \\
=&\int_{o_i,\tilde{o}_{i+1}} \mathbb{E}[\mathbb{E}[w_i^{\pi}(A_i, O_0, H_{i-1}, o_i) \mid S_i, A_i, H_{i-1}]\kappa(A_i, O_i, H_{i-1}, \tilde{o}_{i+1}, o_i)] \text{ by Assumption 3.2} \\
=&\int_{o_i,\tilde{o}_{i+1}} \mathbb{E}[\mathbb{E}[w_i^{\pi}(A_i, O_0, H_{i-1}, o_i) \mid S_i, A_i, H_{i-1}, O_i]\kappa(A_i, O_i, H_{i-1}, \tilde{o}_{i+1}, o_i)] \text{ by Assumption 3.1} \\
=&\int_{o_i,\tilde{o}_{i+1}} \mathbb{E}[\mathbb{E}[w_i^{\pi}(A_i, O_0, H_{i-1}, o_i)\kappa(A_i, O_i, H_{i-1}, \tilde{o}_{i+1}, o_i) \mid S_i, A_i, H_{i-1}, O_i]] \text{ by measurability} \\
=&\int_{o_i,\tilde{o}_{i+1}} \mathbb{E}[w_i^{\pi}(A_i, O_0, H_{i-1}, o_i)\kappa(A_i, O_i, H_{i-1}, \tilde{o}_{i+1}, o_i)] \\
=&\int_{o_i,\tilde{o}_{i+1}} \mathbb{E}[\mathbb{E}[w_i^{\pi}(A_i, O_0, H_{i-1}, o_i)\kappa(A_i, O_i, H_{i-1}, \tilde{o}_{i+1}, o_i) \mid A_t, H_{t-1}, O_0]] \\
=&\int_{o_i,\tilde{o}_{i+1}} \mathbb{E}[w_i^{\pi}(A_i, O_0, H_{i-1}, o_i)\mathbb{E}[\kappa(A_i, O_i, H_{i-1}, \tilde{o}_{i+1}, o_i) \mid A_i, H_{i-1}, O_0]] \\
\leq&\int_{o_i,\tilde{o}_{i+1}} \sqrt{\mathbb{E}[w_i^{\pi}(A_i, H_{i-1}, O_0, o_i)^2]}\sqrt{\mathbb{E}[\{\mathbb{E}[\kappa(A_i, O_i, H_{i-1}, \tilde{o}_{i+1}, o_i) \mid A_i, H_{i-1}, O_0]\}^2]} \\
&\text{(Cauchy-Schwartz inequality)} \\
\leq&\sqrt{\int_{o_i} \mathbb{E}[w_i^{\pi}(A_i, H_{i-1}, O_0, o_i)^2]}\sqrt{\int_{o_i,\tilde{o}_{i+1}} \mathbb{E}[\{\mathbb{E}[\kappa(A_i, O_i, H_{i-1}, \tilde{o}_{i+1}, o_i) \mid A_i, H_{i-1}, O_0]\}^2]} \\
&\text{(Cauchy-Schwartz inequality)} \\
=&C_i^{\pi}\sqrt{\int_{o_i,\tilde{o}_{i+1}} \mathbb{E}[\{\mathbb{E}[\kappa(A_i, O_i, H_{i-1}, \tilde{o}_{i+1}, o_i) \mid A_i, H_{i-1}, O_0]\}^2]} \text{ by definition of concentrability coefficient} \\
\leq&C_i^{\pi}\sqrt{\int_{o_i,\tilde{o}_{i+1}} \mathbb{E}\left[\{\mathbb{E}[U_{i+1}(\tilde{o}_{i+1}, A_i, o_i, H_{i-1}) \mid A_i, H_{i-1}, O_0]\}^2\{\mathbb{E}[\Delta(b_D^{[i]}, b_{D,i})(A_i, O_i, \tilde{o}_{i+1}, o_i) \mid A_i, H_{i-1}, O_0]\}^2\right]} \\
&\text{(Cauchy-Schwartz inequality for conditional expectation)} \\
\leq&C_i^{\pi}\sqrt{\int_{o_i,\tilde{o}_{i+1}} \mathbb{E}\left[(T-i)^2\{\mathbb{E}[\Delta(b_D^{[i]}, b_{D,i})(A_i, O_i, \tilde{o}_{i+1}, o_i) \mid A_i, H_{i-1}, O_0]\}^2\right]} \text{ by Assumption D.3} \\
=&C_i^{\pi}(T-i)\sqrt{\int_{o_i,\tilde{o}_{i+1}} \|p(\tilde{o}_{i+1}, o_i \mid A_i, H_{i-1}, O_0) - \mathbb{E}[b_{D,i}(A_i, O_i, \tilde{o}_{i+1}, o_i) \mid A_i, H_{i-1}, O_0]\|_{\mathcal{L}^2(\mathbb{P}^{\pi^b})}^2} \text{ by Assumption 3.2} \\
=&C_i^{\pi}(T-i)\sqrt{\mathrm{vol}(\mathcal{O}\times\mathcal{O})}\sqrt{\mathcal{L}_D^{[i]}(b_{D,i})} \text{ by Definition A.2.}
\end{aligned} \tag{96}
$$

Therefore, we have

$$
\widetilde{\mathcal{V}}(\pi, \mathbf{b}_R) - \mathcal{V}(\pi, \mathbf{b}_R, \mathbf{b}_D) = \sum_{i=1}^{T-1}(G_{i+1} - G_i) \leq \sum_{i=1}^{T-1} C_i^{\pi}(T-i)\sqrt{\mathrm{vol}(\mathcal{O}\times\mathcal{O})}\sqrt{\mathcal{L}_D^{[i]}(b_{D,i})}. \tag{97}
$$

In addition, by symmetry, it also holds that

$$
\mathcal{V}(\pi, \mathbf{b}_R, \mathbf{b}_D) - \widetilde{\mathcal{V}}(\pi, \mathbf{b}_R) \leq \sum_{i=1}^{T-1} C_i^{\pi}(T-i)\sqrt{\mathrm{vol}(\mathcal{O}\times\mathcal{O})}\sqrt{\mathcal{L}_D^{[i]}(b_{D,i})}. \tag{98}
$$

Combining upper bounds for $|\mathcal{V}(\pi) - \widetilde{\mathcal{V}}(\pi, \mathbf{b}_R)|$ and $|\widetilde{\mathcal{V}}(\pi, \mathbf{b}_R) - \mathcal{V}(\pi, \mathbf{b}_R, \mathbf{b}_D)|$

$$\begin{aligned}
&|\mathcal{V}(\pi) - \mathcal{V}(\pi, \mathbf{b}_R, \mathbf{b}_D)| \\
=&|\mathcal{V}(\pi) - \widetilde{\mathcal{V}}(\pi, \mathbf{b}_R) + \widetilde{\mathcal{V}}(\pi, \mathbf{b}_R) - \mathcal{V}(\pi, \mathbf{b}_R, \mathbf{b}_D)| \\
\leq&|\mathcal{V}(\pi) - \widetilde{\mathcal{V}}(\pi, \mathbf{b}_R)| + |\widetilde{\mathcal{V}}(\pi, \mathbf{b}_R) - \mathcal{V}(\pi, \mathbf{b}_R, \mathbf{b}_D)| \\
\leq&\sum_{t=1}^{T} C_t^{\pi} \sqrt{\mathrm{vol}(\mathcal{R} \times \mathcal{O})} \sqrt{\mathcal{L}_R^{[t]}(b_{R,t})} + \sum_{t=1}^{T-1} C_t^{\pi} (T-t) \sqrt{\mathrm{vol}(\mathcal{O} \times \mathcal{O})} \sqrt{\mathcal{L}_D^{[t]}(b_{D,t})}.
\end{aligned} \tag{99}$$

The proof is completed.

D.4. Proof of Lemma C.2

We aim to show that $\widehat{\mathcal{L}}_R^{[t]}(b_R^{[t]}) - \widehat{\mathcal{L}}_R^{[t]}(\widehat{b}_R^{[t]}) \leq \alpha$ with probability at least $1 - \frac{\delta}{2T}$, and that $\widehat{\mathcal{L}}_D^{[t]}(b_D^{[t]}) - \widehat{\mathcal{L}}_D^{[t]}(\widehat{b}_D^{[t]}) \leq \beta$ with probability at least $1 - \frac{\delta}{2(T-1)}$.

To begin with, we decompose $\widehat{\mathcal{L}}_R^{[t]}(b_R^{[t]}) - \widehat{\mathcal{L}}_R^{[t]}(\widehat{b}_R^{[t]})$ as follows:

$$\begin{aligned}
&\widehat{\mathcal{L}}_R^{[t]}(b_R^{[t]}) - \widehat{\mathcal{L}}_R^{[t]}(\widehat{b}_R^{[t]}) \\
=&\widehat{\mathcal{L}}_R^{[t]}(b_R^{[t]}) - \mathcal{L}_R^{[t]}(b_R^{[t]}) + \mathcal{L}_R^{[t]}(b_R^{[t]}) - \mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) + \mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) - \widehat{\mathcal{L}}_R^{[t]}(\widehat{b}_R^{[t]}) \\
\leq&\widehat{\mathcal{L}}_R^{[t]}(b_R^{[t]}) - \mathcal{L}_R^{[t]}(b_R^{[t]}) + \mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) - \widehat{\mathcal{L}}_R^{[t]}(\widehat{b}_R^{[t]}) \\
\leq&2 \sup_{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t})|
\end{aligned} \tag{100}$$

where the first inequality comes from $\mathcal{L}_R^{[t]}(b_R^{[t]}) - \mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) \leq 0$ as $\mathcal{L}_R^{[t]}(b_R^{[t]}) = 0$ and $\mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) \geq 0$.

It remains to upper bound the term $\sup_{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t})|$.

We employ the following lemma to proceed.

Lemma D.4. *Under Assumptions 3.1, 3.2, 3.3, 4.1, D.16, with probability at least $1 - \delta$, for some $c > 0$, by setting $\lambda_1 = N_1^{-\frac{1}{c_1+1}}$, $\lambda_2 = N_2^{-\frac{\gamma}{\gamma c_2+1}}$, $N_1 = N_2^{\frac{c_1+1}{c_1-1}\frac{\gamma(c_2+1)}{\gamma c_2+1}}$, with probability at least $1 - \delta$, it holds that*

$$\max_{t=1:T} \sup_{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t})| = O_P(M_R \log(T/\delta) N_2^{-\frac{\gamma}{2\gamma+2}})$$

and

$$\max_{t=1:T-1} \sup_{b_{D,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Z}_t}} |\widehat{\mathcal{L}}_D^{[t]}(b_{D,t}) - \mathcal{L}_D^{[t]}(b_{D,t})| = O_P(M_D \log(T/\delta) N_2^{-\frac{\gamma}{2\gamma+2}}).$$

Proof of Lemma D.4. See Appendix D.8 for a proof. □

Next, according to Lemma D.4, and by the definition of $\alpha = c \log(T/\delta) M_R N_2^{-\frac{\gamma}{2\gamma+2}}$, $\beta = c \log(T/\delta) M_D N_2^{-\frac{\gamma}{2\gamma+2}}$, we have

$$\max_{t=1:T} \{\widehat{\mathcal{L}}_R^{[t]}(b_R^{[t]}) - \widehat{\mathcal{L}}_R^{[t]}(\widehat{b}_R^{[t]})\} = O_P(\alpha)$$

and

$$\max_{t=1:T-1} \{\widehat{\mathcal{L}}_D^{[t]}(b_D^{[t]}) - \widehat{\mathcal{L}}_D^{[t]}(\widehat{b}_D^{[t]})\} = O_P(\beta).$$

The proof is done.

D.5. Proof of Lemma C.3

We aim to show that $\mathcal{L}_R^{[t]}(b_{R,t})$ can be upper bounded uniformly for any $b_{R,t} \in \text{conf}_R(\alpha)$ and uniformly for all $t = 1, .., T$. Similarly, we need to show $\mathcal{L}_D^{[t]}(b_{D,t})$ can be upper bounded uniformly for any $b_{D,t} \in \text{conf}_D(\beta)$ and uniformly for all $t = 1, .., T-1$.

We first present the following decomposition.

$$\begin{aligned}
&\mathcal{L}_R^{[t]}(b_{R,t}) \\
=&\mathcal{L}_R^{[t]}(b_{R,t}) - \widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) + \widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \widehat{\mathcal{L}}_R^{[t]}(\widehat{b}_R^{[t]}) + \widehat{\mathcal{L}}_R^{[t]}(\widehat{b}_R^{[t]}) - \mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) + \mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) \\
\leq& 2\sup_{b_{R,t}} |\mathcal{L}_R^{[t]}(b_{R,t}) - \widehat{\mathcal{L}}_R^{[t]}(b_{R,t})| + \alpha + \mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}).
\end{aligned} \tag{101}$$

Similarly, we have

$$\begin{aligned}
&\mathcal{L}_D^{[t]}(b_{D,t}) \\
\leq& 2\sup_{b_{D,t}} |\mathcal{L}_D^{[t]}(b_{D,t}) - \widehat{\mathcal{L}}_D^{[t]}(b_{D,t})| + \beta + \mathcal{L}_D^{[t]}(\widehat{b}_D^{[t]}).
\end{aligned} \tag{102}$$

We need Lemma D.4 and the following lemma to proceed.

Lemma D.5. *Under Assumptions 3.1, 3.2, 3.3, 4.1, D.16, with probability at least $1-\delta$, for some $c > 0$, by setting $\lambda_1 = N_1^{-\frac{1}{c_1+1}}$, $\lambda_2 = N_2^{-\frac{\gamma}{\gamma c_2+1}}$, $N_1 = N_2^{\frac{c_1+1}{c_1-1}\frac{\gamma(c_2+1)}{\gamma c_2+1}}$, and setting*

$$\alpha = c\log(T/\delta)M_R N_2^{-\frac{\gamma}{2\gamma+2}}, \beta = c\log(T/\delta)M_D N_2^{-\frac{\gamma}{2\gamma+2}},$$

it holds that

$$\max_{t=1:T} \mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) = O_P(\log(T/\delta)N_2^{-\frac{\gamma c_2}{\gamma c_2+1}})$$

and

$$\max_{t=1:T-1} \mathcal{L}_D^{[t]}(\widehat{b}_D^{[t]}) = O_P(\log(T/\delta)N_2^{-\frac{\gamma c_2}{\gamma c_2+1}}).$$

Here γ are defined in Assumption 4.1.

Proof of Lemma D.5. See Appendix D.7 for a proof. □

In particular, according to Lemma D.4, we have

$$\max_{t=1:T} \sup_{b_{R,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t})| = O_P(M_R\log(T/\delta)N_2^{-\frac{\gamma}{2\gamma+2}})$$

and

$$\max_{t=1:T-1} \sup_{b_{D,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Z}_t}} |\widehat{\mathcal{L}}_D^{[t]}(b_{D,t}) - \mathcal{L}_D^{[t]}(b_{D,t})| = O_P(M_D\log(T/\delta)N_2^{-\frac{\gamma}{2\gamma+2}}).$$

We can see that $\max_{t=1:T}\sup_{b_{R,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t})|$ and $\max_{t=1:T-1}\sup_{b_{D,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Z}_t}} |\widehat{\mathcal{L}}_D^{[t]}(b_{D,t}) - \mathcal{L}_D^{[t]}(b_{D,t})|$ are the dominating terms, which are of the order α, β respectively.

Therefore, we have $\max_{t=1:T}\sup_{b_{R,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} \mathcal{L}_R^{[t]}(b_{R,t}) = O_P(\alpha)$ and $\max_{t=1:T-1}\sup_{b_{D,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Z}_t}} \mathcal{L}_D^{[t]}(b_{D,t}) = O_P(\beta)$. The proof is completed by taking a square root at both sides.

D.6. Proof of Lemma D.1

We need to show

$$\sum_{a_t,a_{t-1},\ldots,a_1}\int_{o_t,o_{t-1},\ldots,o_1}\prod_{j=1}^{t}\pi_j(a_j\mid o_j,h_{j-1})\int_{s_t,s_{t-1},\ldots,s_1}\int_{\tilde{o}_t}g(a_t,\tilde{o}_t,h_{t-1},o_t)p(\tilde{o}_t\mid s_t)\prod_{j=1}^{t-1}p(s_{j+1},o_j\mid s_j,a_j)p(s_1)$$
$$=\int_{o_t}\mathbb{E}\left[\frac{p^{\pi}\left(S_t,H_{t-1}\right)\pi_t\left(A_t\mid o_t,H_{t-1}\right)}{p^{\pi^b}\left(S_t,H_{t-1}\right)\pi_t^b\left(A_t\mid S_t\right)}g\left(A_t,O_t,H_{t-1},o_t\right)\right]. \tag{103}$$

For clarity, we simply use $\int_{a_t}$ to denote $\sum_{a_t}$ for every t. Then we do a direct calculation in the following

$$\begin{aligned}
&\int_{o_t}\mathbb{E}\left[\frac{p^{\pi}\left(S_t,H_{t-1}\right)\pi_t\left(A_t\mid o_t,H_{t-1}\right)}{p^{\pi^b}\left(S_t,H_{t-1}\right)\pi_t^b\left(A_t\mid S_t\right)}g\left(A_t,O_t,H_{t-1},o_t\right)\right]\\
&=\int_{a_t,o_t,\tilde{o}_t,s_t,h_{t-1}}\frac{p^{\pi}(s_t,h_{t-1})\pi_t(a_t\mid o_t,h_{t-1})}{p^{\pi^b}(s_t,h_{t-1})\pi_t^b(a_t\mid s_t)}g(a_t,\tilde{o}_t,h_{t-1},o_t)p^{\pi^b}(a_t,\tilde{o}_t,s_t,h_{t-1})\\
&=\int_{a_t,o_t,\tilde{o}_t,s_t,h_{t-1}}\frac{p^{\pi}(s_t,h_{t-1})\pi_t(a_t\mid o_t,h_{t-1})}{p^{\pi^b}(s_t,h_{t-1})\pi_t^b(a_t\mid s_t)}g(a_t,\tilde{o}_t,h_{t-1},o_t)p^{\pi^b}(a_t,\tilde{o}_t\mid s_t,h_{t-1})p^{\pi^b}(s_t,h_{t-1})\\
&=\int_{a_t,o_t,\tilde{o}_t,s_t,h_{t-1}}\frac{p^{\pi}(s_t,h_{t-1})\pi_t(a_t\mid o_t,h_{t-1})}{p^{\pi^b}(s_t,h_{t-1})\pi_t^b(a_t\mid s_t)}g(a_t,\tilde{o}_t,h_{t-1},o_t)p^{\pi^b}(a_t,\tilde{o}_t\mid s_t)p^{\pi^b}(s_t,h_{t-1})\\
&=\int_{a_t,o_t,\tilde{o}_t,s_t,h_{t-1}}\frac{p^{\pi}(s_t,h_{t-1})\pi_t(a_t\mid o_t,h_{t-1})}{p^{\pi^b}(s_t,h_{t-1})\pi_t^b(a_t\mid s_t)}g(a_t,\tilde{o}_t,h_{t-1},o_t)\pi_t^b(a_t\mid s_t)p(\tilde{o}_t\mid s_t)p^{\pi^b}(s_t,h_{t-1})\text{ by }A_t\perp\!\!\!\perp O_t\mid S_t\\
&=\int_{a_t,o_t,\tilde{o}_t,s_t,h_{t-1}}\pi_t(a_t\mid o_t,h_{t-1})g(a_t,\tilde{o}_t,h_{t-1},o_t)p(\tilde{o}_t\mid s_t)p^{\pi}(s_t,h_{t-1}).
\end{aligned}\tag{104}$$

We then provide an expression of $p^{\pi}(s_t,h_{t-1})$ in the following:

$$\begin{aligned}
&p^{\pi}(s_t,h_{t-1})\\
&=\int_{s_{t-1}}p^{\pi}(s_t,s_{t-1},h_{t-1})\\
&=\int_{s_{t-1}}p^{\pi}(s_t\mid s_{t-1},h_{t-1})p^{\pi}(s_{t-1},h_{t-1})\\
&=\int_{s_{t-1}}p(s_t\mid s_{t-1},a_{t-1})p^{\pi}(s_{t-1},h_{t-1})\text{ by }S_t\perp\!\!\!\perp(H_{t-2},O_{t-1})\mid(S_t,A_t)\\
&=\int_{s_{t-1}}p(s_t\mid s_{t-1},a_{t-1})p^{\pi}(s_{t-1},a_{t-1},o_{t-1},h_{t-2})\\
&=\int_{s_{t-1}}p(s_t\mid s_{t-1},a_{t-1})p^{\pi}(a_{t-1}\mid s_{t-1},o_{t-1},h_{t-2})p^{\pi}(o_{t-1}\mid s_{t-1},h_{t-2})p^{\pi}(s_{t-1},h_{t-2})\\
&=\int_{s_{t-1}}p(s_t\mid s_{t-1},a_{t-1})\pi_{t-1}(a_{t-1}\mid o_{t-1},h_{t-2})p(o_{t-1}\mid s_{t-1})p^{\pi}(s_{t-1},h_{t-2})\\
&=\int_{s_{t-1}}p(s_t\mid s_{t-1},a_{t-1})\pi_{t-1}(a_{t-1}\mid o_{t-1},h_{t-2})p(o_{t-1}\mid s_{t-1},a_{t-1})p^{\pi}(s_{t-1},h_{t-2})\\
&=\int_{s_{t-1}}p(s_t,o_{t-1}\mid s_{t-1},a_{t-1})\pi_{t-1}(a_{t-1}\mid o_{t-1},h_{t-2})p^{\pi}(s_{t-1},h_{t-2})\\
&=\cdots\\
&=\int_{s_{t-1},s_{t-2},\ldots,s_1}\prod_{j=1}^{t-1}p(s_{j+1},o_j\mid s_j,a_j)\prod_{j=1}^{t-1}\pi_j(a_j\mid o_j,h_{j-1})p(s_1).
\end{aligned}\tag{105}$$

By plugging (105) into the last equality of (104), we have

$$\begin{aligned}
&\int_{o_t} \mathbb{E}\left[\frac{p^{\pi}\left(S_t, H_{t-1}\right) \pi_t\left(A_t \mid o_t, H_{t-1}\right)}{p^{\pi^b}\left(S_t, H_{t-1}\right) \pi_t^b\left(A_t \mid S_t\right)} g\left(A_t, O_t, H_{t-1}, o_t\right)\right] \\
=&\int_{a_t, o_t, \tilde{o}_t, s_t, h_{t-1}} \pi_t(a_t \mid o_t, h_{t-1}) g(a_t, \tilde{o}_t, h_{t-1}, o_t) p(\tilde{o}_t \mid s_t) p^{\pi}(s_t, h_{t-1}) \\
=&\int_{a_t, o_t, \tilde{o}_t, s_t, h_{t-1}} \pi_t(a_t \mid o_t, h_{t-1}) g(a_t, \tilde{o}_t, h_{t-1}, o_t) p(\tilde{o}_t \mid s_t) \int_{s_{t-1}, s_{t-2}, \ldots, s_1} \\
&\prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) \prod_{j=1}^{t-1} \pi_j(a_j \mid o_j, h_{j-1}) p(s_1) \\
=&\int_{h_t} \prod_{j=1}^{t} \pi_j(a_j \mid o_j, h_{j-1}) \int_{s_t, s_{t-1}, s_{t-2}, \ldots, s_1} \int_{\tilde{o}_t} g(a_t, \tilde{o}_t, h_{t-1}, o_t) p(\tilde{o}_t \mid s_t) \prod_{j=1}^{t-1} p(s_{j+1}, o_j \mid s_j, a_j) p(s_1).
\end{aligned} \tag{106}$$

The proof is done.

D.7. Proof of Lemma D.5

We prove finite-sample upper bounds on $\mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]})$ and $\mathcal{L}_D^{[t]}(\widehat{b}_D^{[t]})$ in this proof. Most of the proof for this lemma is adapted from Mastouri et al. (2021); Singh et al. (2019); Szabó et al. (2016); Caponnetto & De Vito (2007), excepted that we need to deal with $\widehat{p}(y_t \mid x_t)$, $\widehat{p}(z_t \mid x_t)$ in this work rather than the true observed $p(y_t \mid x_t)$, $p(z_t \mid x_t)$. To begin with, we focus on $\mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]})$. We are going to analyze the stage 1 error at first, and then analyze the stage 2 error.

We first introduce some notations for the sake of convenience. As $\mathcal{H}_{\mathcal{X}} \otimes \mathcal{H}_{\mathcal{Y}}$ is isometrically isomorphic to $\mathcal{H}_{\mathcal{XY}}$, we use their features interchangeably, i.e. $\phi(x, y) = \phi(x) \otimes \phi(y)$. $k(\cdot, \cdot)$ is a general notation for a kernel function, and $\phi(\cdot)$ denotes RKHS feature maps. To simplify notation, the argument of the kernel/feature map identifies it: for instance, $k(x, \cdot)$ and $\phi(x)$ denote the respective kernel and feature map on $\mathcal{X}$. We denote $K_{x\tilde{x}} := k(x, \tilde{x})$.

For any Hilbert space $\mathcal{F}$, we denote $\mathcal{L}(\mathcal{F})$ the space of bounded linear operators from $\mathcal{F}$ to itself. For any Hilbert space $\mathcal{G}$, we denote by $\mathcal{L}^2(\mathcal{F}, \mathcal{G})$ the space of Hilbert-Schmidt operators from $\mathcal{F}$ to $\mathcal{G}$. We denote by $L^2(\mathcal{F}, \mathbb{P}^{\pi^b})$ the space of square integrable functions on $\mathcal{F}$ with respect to measure $\mathbb{P}^{\pi^b}$.

We analyze the stage 1 estimation at first.

Stage 1. At stage 1, we learn the conditional mean embedding of the conditional distribution $p(w_t \mid x_t)$. There exist some works that have studied the finite-sample convergence of the conditional mean embedding. Here, we directly adopt the theoretical results (including required assumptions, definitions, theorems, etc) from existing works Singh et al. (2019); Mastouri et al. (2021) regarding the analysis of stage 1 error.

The optimal $C_{W_t|X_t}$ minimizes the expected discrepancy:

$$C_{W_t|X_t} = \underset{C \in \mathcal{L}^2\left(\mathcal{H}_{\mathcal{X}_t}, \mathcal{H}_{\mathcal{W}_t}\right)}{\operatorname{argmin}} E_t(C), \text{ where } E_t(C) = \mathbb{E}\|\phi(W_t) - C\phi(X_t)\|^2_{\mathcal{H}_{\mathcal{W}_t}} \tag{107}$$

According to Song et al. (2009; 2013), it suffices to solve a vector-valued regression in order to learn $C_{W_t|X_t}$. The search space in the regression problem is the vector-valued RKHS $\mathcal{H}_{\Gamma_t}$ of operators mapping $\mathcal{H}_{\mathcal{X}_t}$ to $\mathcal{H}_{\mathcal{W}_t}$. See also a review of the kernel conditional mean embedding Muandet et al. (2017). In particular, $\mathcal{H}_{\mathcal{X}_t} \otimes \mathcal{H}_{\mathcal{W}_t}$ is isomorphic to $\mathcal{L}^2\left(\mathcal{H}_{\mathcal{X}_t}, \mathcal{H}_{\mathcal{W}_t}\right)$. Therefore, by choosing the vector-valued kernel Γ_t with feature map : $(x_t, w_t) \mapsto [\phi(x_t) \otimes \phi(w_t)] := \phi(x_t)\langle\phi(w_t), \cdot\rangle_{\mathcal{H}_{\mathcal{W}_t}}$, we have $\mathcal{H}_{\Gamma_t} = \mathcal{L}^2\left(\mathcal{H}_{\mathcal{X}_t}, \mathcal{H}_{\mathcal{W}_t}\right)$ and they share the same norm. We denote by $L^2(\mathcal{X}_t, \mathbb{P}^{\pi^b}_{\mathcal{X}_t})$ the space of square integrable functions from $\mathcal{X}_t$ to $\mathcal{W}_t$ with respect to measure $\mathbb{P}^{\pi^b}_{\mathcal{X}_t}$, where $\mathbb{P}^{\pi^b}_{\mathcal{X}_t}$ is the restriction of $\mathbb{P}^{\pi^b}$ to $\mathcal{X}_t$.

We drop the subscript with respect to t in the following if it does not cause confusion. Also, we adopt the same notations and results directly from Mastouri et al. (2021), and they are only used in this proof.

The following assumptions and definitions are needed.

Assumption D.6. For each $t = 1, ..., T$, $\mathcal{X}_t, \mathcal{Y}_t, \mathcal{W}_t, \mathcal{Z}_t$ are measurable, separable Polish spaces.

Assumption D.7. (i) $k(w, \cdot)$ is a characteristic kernel. (ii) $k(y, \cdot), k(x, \cdot), k(w, \cdot)$ and $k(z, \cdot)$ are continuous, bounded by $\kappa > 0$, and their feature maps are measurable.

The kernel mean embedding of any probability distribution is injective if a characteristic kernel is used (Sriperumbudur et al., 2011); this guarantees that a probability distribution can be uniquely represented in an RKHS.

Assumption D.8. For each $t = 1, ..., T$, assume that $C_{W_t|X_t} \in \mathcal{H}_{\Gamma_t}$, i.e. $C_{W_t|X_t} = \text{argmin}_{C \in \mathcal{H}_{\Gamma_t}} E_t(C)$

Definition D.9 (Kernel Integral operator for Stage 1)**.** Define the integral operator:

$$S_1 : L^2\left(\mathcal{X}, \mathbb{P}^{\pi^b}_{\mathcal{X}}\right) \longrightarrow \mathcal{H}_{\mathcal{X}}$$
$$g \longmapsto \int \phi(x) g(x) d\mathbb{P}^{\pi^b}_{\mathcal{X}}(x).$$

The uncentered covariance operator is defined by $T_1 = S_1 \circ S_1^*$, where S_1^* is the adjoint of S_1.

Assumption D.10. Fix $\gamma_1 < \infty$. For given $c_1 \in (1, 2]$, we assume that $\exists G_1 \in \mathcal{H}_\Gamma$ s.t. $C_{W|X} = T_1^{\frac{c_1-1}{2}} \circ G_1$ and $\|G_1\|^2_{\mathcal{H}_\Gamma} \leq \gamma_1$.

The following theorem provides a closed-form solution to the ERM in stage 1.

Theorem D.11 (Singh et al. (2019), Theorem 1)**.** *For any $\lambda_1 > 0$, the solution of (10) exists, is unique, and is given by:*

$$\widehat{C}_{W|X} = (\boldsymbol{T}_1 + \lambda_1)^{-1} g_1, \text{ where } \boldsymbol{T}_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} \phi(x_i) \otimes \phi(x_i),$$
$$\text{and } g_1 = \frac{1}{N_1} \sum_{i=1}^{N_1} \phi(x_i) \otimes \phi(w_i)$$

and for any $x \in \mathcal{X}$, we have $\widehat{\mu}_{W|x} = \widehat{C}_{W|X}\phi(x)$.

The next theorem provides a finite-sample upper bound on the estimation error.

Theorem D.12 (Finite-sample upper bound at stage 1, Mastouri et al. (2021), Theorem 5)**.** *Suppose Assumptions D.6, D.7, D.8 and D.10 hold. Define λ_1 as:*

$$\lambda_1 = \left(\frac{8\kappa \left(\kappa + \kappa \left\|C_{W|X}\right\|_{\mathcal{H}_\Gamma}\right) \ln(2/\delta)}{\sqrt{N_1 \gamma_1 (c_1 - 1)}} \right)^{\frac{2}{c_1+1}}$$

Then, for any $x \in \mathcal{X}$ and any $\delta \in (0, 1)$, the following holds with probability $1 - \delta$:

$$\left\|\widehat{\mu}_{W|x} - \mu_{W|x}\right\|_{\mathcal{H}_\mathcal{W}} \leq \kappa r_C(\delta, N_1, c_1) =: \kappa \frac{\sqrt{\gamma_1}(c_1 + 1)}{4^{\frac{1}{c_1+1}}} \left(\frac{4\kappa \left(\kappa + \kappa \left\|C_{W|X}\right\|_{\mathcal{H}_\Gamma}\right) \ln(2/\delta)}{\sqrt{N_1 \gamma_1}(c_1 - 1)} \right)^{\frac{c_1-1}{c_1+1}}$$

where $\widehat{\mu}_{W|X} = \widehat{C}_{W|X}\phi(x)$ and $\widehat{C}_{W|X}$ is the solution of (10).

The proof of Theorem D.12 is omitted in this paper. Readers can refer to Mastouri et al. (2021); Singh et al. (2019) for a detailed proof.

Corollary D.13. *Under the same conditions from Theorem D.12, for any $x_t \in \mathcal{X}_t$ and any $\delta \in (0, 1)$, the following holds uniformly for all $t = 1, ..., T$ with probability $1 - \delta$:*

$$\left\|\widehat{\mu}_{W_t|x_t} - \mu_{W_t|x_t}\right\|_{\mathcal{H}_{\mathcal{W}_t}} \leq \kappa \frac{\sqrt{\gamma_1}(c_1 + 1)}{4^{\frac{1}{c_1+1}}} \left(\frac{4\kappa \left(\kappa + \kappa \left\|C_{W_t|X_t}\right\|_{\mathcal{H}_{\Gamma_t}}\right) \ln(2T/\delta)}{\sqrt{N_1 \gamma_1}(c_1 - 1)} \right)^{\frac{c_1-1}{c_1+1}}$$

Proof of Corollary D.13. Let $\delta := \frac{\delta'}{T}$ in Theorem D.12 and apply a union bound argument. □

Stage 2.

We then analyze the stage 2 error where the output from stage 1 is used as a plug-in estimator. Most of the proof for the stage 2 analysis can be adapted from Singh et al. (2019); Mastouri et al. (2021), except that we need to deal with $\widehat{p}(w_t \mid x_t)$ rather than $p(w_t \mid x_t)$.

The optimal $b^*_{R,t}$ minimizes the expected discrepancy:

$$b^*_{R,t} = \underset{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}}{\operatorname{argmin}} \mathcal{L}^{[t]}_R(b_{R,t}) \text{ where} \tag{108}$$

$$\mathcal{L}^{[t]}_R(b_{R,t}) = \mathbb{E}_{X_t \sim \mathbb{P}^{\pi^b}, Y_t \sim \mathrm{Unif}(\mathcal{Y})} \left(\left\langle \mu_{W_t|X_t} \otimes \phi\left(Y_t\right), b_{R,t} \right\rangle_{\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} - p\left(Y_t \mid X_t\right) \right)^2. \tag{109}$$

Throughout the proof of stage 2 error, we use the notation $\widetilde{\mathbb{P}}$ to denote the probability measure, in which both W_t, X_t follows $\mathbb{P}^{\pi^b}$ while Y_t follows $\mathrm{unif}(\mathcal{Y})$.

Similarly to Stage 1, the problem of learning $b^*_{R,t}$ is transformed into a ridge regression. We list some needed assumptions and definitions as well in the following.

Assumption D.14. For each $t = 1, ..., T$, we assume that $b^{[t]}_R = b^*_{R,t}$, i.e. the minimization problem $\underset{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}}{\operatorname{argmin}} \mathcal{L}^{[t]}_R(b_{R,t})$ is achievable with $\min_{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} \mathcal{L}^{[t]}_R(b_{R,t}) = 0$.

Definition D.15. (Kernel integral operator for Stage 2). Define the integral operator :

$$\begin{aligned} S_2 : \mathcal{H}_{\mathcal{W}\mathcal{Y}} &\longrightarrow \mathcal{H}_{\mathcal{W}\mathcal{Y}} \\ b &\longmapsto \int \left[\mu_{W|x} \otimes \phi(y)\right] b \left[\phi(y) \otimes \mu_{W|x}\right] d\widetilde{\mathbb{P}}_{\mathcal{H}_{\mathcal{W}} \times \mathcal{X} \times \mathcal{Y}} \left(\mu_{W|x}, y\right). \end{aligned}$$

The uncentered covariance operator is defined by $T_2 = S_2 \circ S_2^*$, where S_2^* is the adjoint of S_2.

Assumption D.16. Fix $\gamma_2 < \infty$. For given $c_2 \in (1, 2]$, we assume that $\widetilde{\mathbb{P}}$ belongs to a prior class of functions $\mathcal{P}\left(\gamma_2, b_2, c_2\right)$ such that:
(a) A range space assumption is satisfied : $\exists G_2 \in \mathcal{H}_{\mathcal{W}\mathcal{Y}}$ s.t. $b^*_R = T_2^{\frac{c_2-1}{2}} \circ G_2$, $b^*_D = T_2^{\frac{c_2-1}{2}} \circ G_2$ and $\|G_2\|_{\mathcal{H}_{\mathcal{W}\mathcal{Y}}} \leq \gamma_2$
(b) The eigenvalues $(l_k)_{k \in \mathbb{N}^*}$ of T_2 satisfy $\alpha_2 \leq l_k k^{b_2} \leq \beta_2$ for $b_2 > 1, \alpha_2, \beta_2 > 0$.
(c) The conditional density function $p(y_t \mid x_t)$ is uniformly bounded by a constant m for each $t = 1, ..., T$.
(d) Without loss of generality, $\mathcal{H}_{\mathcal{W}} \otimes \mathcal{H}_{\mathcal{Y}}$ is a $\| \cdot \|_{\mathcal{H}_{\mathcal{W}} \otimes \mathcal{H}_{\mathcal{Y}}}$ normed constrained space with $\|b\|_{\mathcal{H}_{\mathcal{W}} \otimes \mathcal{H}_{\mathcal{Y}}} \leq 1$ for all $b \in \mathcal{H}_{\mathcal{W}} \otimes \mathcal{H}_{\mathcal{Y}}$.
(e) There exist $M_R > 0$, $M_D > 0$ such that $|b_{R,t}|_\infty \leq M_R$, $|b_{D,t}|_\infty \leq M_D$ for every $b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}$, $\forall t = 1, ..., T$ $b_{D,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Z}_t}$, $\forall t = 1, ..., T-1$.

Next, we introduce a notation $\widetilde{b}^{[t]}_R$ which is the minimizer of the empirical risk of stage 2, when plugging in the true $\mu_{W_t|x_t}$ and the true $p(y_t \mid x_t)$ instead of their estimates:

$$\begin{aligned} \widetilde{b}^{[t]}_R &= \underset{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \bigotimes \mathcal{H}_{\mathcal{Y}_t}}{\operatorname{argmin}} \widetilde{\mathcal{L}}^{[t]}_R(b_{R,t}), \text{ where} \\ \widetilde{\mathcal{L}}^{[t]}_R(b_{R,t}) &= \frac{1}{N_2} \sum_{n=1}^{N_2} \left(\left\langle \mu_{W_t|x'_{t,n}} \otimes \phi\left(y''_{t,n}\right), b_{R,t} \right\rangle_{\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} - p\left(y''_{t,n} \mid x'_{t,n}\right) \right)^2 + \lambda_2 \|b_{R,t}\|^2_{\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}}. \end{aligned} \tag{110}$$

Similarly to $\widehat{b}^{[t]}_R$, it has a closed form solution given below (see Grunewalder et al. (2012), section D.1).

Theorem D.17. *For any $\lambda_2 > 0$, the solutions of (110) (with dropped subscript t), exists, is unique, and is given by $(\boldsymbol{T}_2 + \lambda_2)^{-1} g_2$, where*

$$\begin{aligned} \boldsymbol{T}_2 &= \frac{1}{N_2} \sum_{n=1}^{N_2} \left[\mu_{W|x_n'} \otimes \phi\left(y_n''\right)\right] \otimes \left[\mu_{W|x_n'} \otimes \phi\left(y_n''\right)\right], \\ g_2 &= \frac{1}{M} \sum_{n=1}^{N_2} \left[\mu_{W|x_n'} \otimes \phi\left(y_n''\right)\right] p(y_n'' \mid x_n'). \end{aligned} \tag{111}$$

Define also $b_R^{\lambda_2}$ as the minimizer of the population version of (110):

$$\begin{aligned} b_R^{\lambda_2} &= \underset{b_R \in \mathcal{H}_{\mathcal{W}} \bigotimes \mathcal{H}_{\mathcal{Y}}}{\operatorname{argmin}} \mathcal{L}_R^{\lambda_2}(b_R), \text{ where} \\ \mathcal{L}_R^{\lambda_2}(b_R) &= \mathbb{E}_{X \sim \mathbb{P}^{\pi^b}, Y \sim \operatorname{unif}(\mathcal{Y})} \left(\left\langle \mu_{W|X} \otimes \phi(Y), b_R \right\rangle_{\mathcal{H}_{\mathcal{W}} \otimes \mathcal{H}_{\mathcal{Y}}} - p(Y \mid X)\right)^2 + \lambda_2 \|b_R\|^2_{\mathcal{H}_{\mathcal{W}} \otimes \mathcal{H}_{\mathcal{Y}}} \end{aligned} \tag{112}$$

where we dropped the subscript t if no confusion is caused.

Then the upper bound on the $\mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) - \mathcal{L}_R^{[t]}(b_R^{[t]})$ can be bounded by several terms that involve the stage 1 error, stage 2 error, and approximation error.

Lemma D.18 (Mastouri et al. (2021), Proposition 5)**.** *The following inequality holds.*

$$\mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) - \mathcal{L}_R^{[t]}(b_R^{[t]}) \leq 5\left[S_{-1} + S_0 + \mathcal{A}(\lambda_2) + S_1 + S_2\right]$$

where

$$\begin{aligned} S_{-1} &= \left\| \sqrt{T_2} \circ \left(\widehat{\boldsymbol{T}}_{\boldsymbol{2}} + \lambda_2\right)^{-1} (\widehat{g}_2 - g_2) \right\|^2_{\mathcal{H}_{\mathcal{W}\mathcal{Y}}}, \quad S_0 = \left\| \sqrt{T_2} \circ \left(\widehat{\boldsymbol{T}}_{\boldsymbol{2}} + \lambda_2\right)^{-1} \circ \left(\boldsymbol{T}_{\boldsymbol{2}} - \widehat{\boldsymbol{T}}_{\boldsymbol{2}}\right) \widetilde{b}_R^{[t]} \right\|^2_{\mathcal{H}_{\mathcal{W}\mathcal{Y}}} \\ S_1 &= \left\| \sqrt{T_2} \circ (\boldsymbol{T}_{\boldsymbol{2}} + \lambda_2)^{-1} \left(g_2 - \boldsymbol{T}_2 b_R^{[t]}\right) \right\|^2_{\mathcal{H}_{\mathcal{W}\mathcal{Y}}}, \quad S_2 = \left\| \sqrt{T_2} \circ (\boldsymbol{T}_{\boldsymbol{2}} + \lambda_2)^{-1} \circ (T_2 - \boldsymbol{T}_{\boldsymbol{2}}) \left(b_R^{[t],\lambda_2} - b_R^{[t]}\right) \right\|^2_{\mathcal{H}_{\mathcal{W}\mathcal{Y}}} \end{aligned}$$

and the residual $\mathcal{A}(\lambda_2) = \left\| \sqrt{T_2} \left(b_R^{[t],\lambda_2} - b_R^{[t]}\right) \right\|^2_{\mathcal{H}_{\mathcal{W}\mathcal{Y}}}$.

Proof. According to Proposition 2 in Vito & Caponnetto (2005), the excess risk can be decomposed as

$$\begin{aligned} \mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) - \mathcal{L}_R^{[t]}(b_R^{[t]}) &= \left\| \sqrt{T_2} \left(\widehat{b}_R^{[t]} - b_R^{[t]}\right) \right\|^2_{\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} \\ &= \left\| \sqrt{T_2} \left(\widehat{b}_R^{[t]} - \widetilde{b}_R^{[t]} + \widetilde{b}_R^{[t]} - b_R^{[t],\lambda_2} + b_R^{[t],\lambda_2} - b_R^{[t]}\right) \right\|^2_{\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}}. \end{aligned} \tag{113}$$

Then, readers can refer to the proof of Proposition 5 in Mastouri et al. (2021). □

Intuitively, S_{-1}, S_0 quantify the estimation error at stage 1; S_1, S_2 quantify the estimation error at stage 2; $\mathcal{A}(\lambda_2)$ quantifies the bias/approximation error from the regularized regression.

The upper bounds on S_1, S_2 and $\mathcal{A}(\lambda_2)$ can be directly adapted from Vito & Caponnetto (2005); Caponnetto & De Vito (2007). It is because S_1, S_2, $\mathcal{A}(\lambda_2)$ have replaced the estimates of conditional mean embedding and conditional distribution from stage 1 to the true ones. In this way, S_1, S_2, $\mathcal{A}(\lambda_2)$ can be viewed as the errors from a regularized least square problem, which has been studied by Vito & Caponnetto (2005); Caponnetto & De Vito (2007). The following two lemmas provide upper bounds on S_1, S_2, $\mathcal{A}(\lambda_2)$.

Lemma D.19 (Mastouri et al. (2021), Proposition 6)**.** *Suppose Assumption D.16 holds. Then, the residual $\mathcal{A}(\lambda_2)$, the reconstruction error $\mathcal{B}(\lambda_2)$, and the effective dimension $\mathcal{N}(\lambda_2)$ are defined and bounded as follows:*

$$\mathcal{A}(\lambda_2) = \left\| \sqrt{T_2} \left(b_R^{[t],\lambda_2} - b_R^{[t]}\right) \right\|^2_{\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} \leq \gamma_2 \lambda_2^{c_2}, \quad \mathcal{B}(\lambda_2) = \left\| b_R^{[t],\lambda_2} - b_R^{[t]} \right\|^2_{\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} \leq \gamma_2 \lambda_2^{c_2 - 1}$$

$$\mathcal{N}(\lambda_2) = \operatorname{Tr}\left[(T_2 + \lambda_2)^{-1} \circ T_2\right] \leq \beta_2^{\frac{1}{b_2}} \frac{\pi/b_2}{\sin(\pi/b_2)} \lambda_2^{-\frac{1}{b_2}}$$

Lemma D.20 (Theorem 4 of Caponnetto & De Vito (2007), Proposition 7 of Mastouri et al. (2021))**.** *Assume Assumption D.14 and Assumption D.16 hold. Assume also that* $\lambda_2 \leq \|T_2\|_{\mathcal{L}(\mathcal{H}_{\mathcal{W}\mathcal{Y}})}$ *and* $N_2 \geq \frac{2C_\epsilon \mathcal{N}(\lambda_2)}{\lambda_2}$*. Then, we can bound* S_1 *and* S_2 *from Lemma D.18 as follows w.p.* $1-2\epsilon/3$ *:*

$$S_1 \leq 32\ln^2(6/\epsilon)\left[\frac{\left(m+\left\|b_R^{[t]}\right\|_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}}\right)^2(4+N_2\lambda_2\mathcal{N}(\lambda_2))}{N_2^2\lambda_2}\right], \quad S_2 \leq 8\ln^2(6/\epsilon)\left[\frac{4\mathcal{B}(\lambda_2)+N_2\mathcal{A}(\lambda_2)}{N_2^2\lambda_2}\right].$$

Next, we focus on the stage 1 errors S_{-1} and S_0. The following lemmas are needed.

Lemma D.21 (Proposition 8 of Mastouri et al. (2021))**.** *Assume the assumptions of Theorem D.12 hold and define* λ_1 *accordingly. Suppose also that Assumption D.14 and Assumption D.16 hold. Then, w.p.* $1-\delta$ *:*

$$\left\|\boldsymbol{T}_2-\widehat{\boldsymbol{T}}_2\right\|^2_{\mathcal{L}(\mathcal{H}_{\mathcal{W}}\otimes\mathcal{H}_{\mathcal{Y}})} \leq 4\kappa^6 r_C(\delta,N_1,c_1)^2$$

Lemma D.22 (MLE guarantee.)**.** *Given a set of models* $\mathcal{M}=\{P:\mathcal{X}\to\Delta(\mathcal{Y})\}$ *with* $P^\star\in\mathcal{M}$*, and a dataset* $\mathcal{D}=\{x_i,y_i\}_{i=1}^{N_1}$ *following* $\mathbb{P}^{\pi^b}$*, let* $\widehat{P}_{MLE}$ *be*

$$\widehat{P}_{\mathrm{MLE}}=\underset{P\in\mathcal{M}}{\arg\min}\sum_{i=1}^{N_1}-\ln P(y_i\mid x_i).$$

With probability at least $1-\delta$*, we have:*

$$\mathbb{E}_{x\sim\mathbb{P}^{\pi^b}}\mathrm{TV}\left(\widehat{P}_{\mathrm{MLE}}(\cdot\mid x),P^\star(\cdot\mid x)\right)^2 \lesssim \frac{\ln(|\mathcal{M}|/\delta)}{N_1}$$

Proof. See Agarwal et al. (2020)[Section E] for a proof. □

Lemma D.23. *Assume the assumptions of Theorem D.12 and assumptions of Lemma D.22 hold and define* λ_1 *accordingly. Suppose also that Assumption D.14 and Assumption D.16 hold. Then, w.p.* $1-3\delta$ *:*

$$\|\widehat{g}_2-g_2\|^2_{\mathcal{H}_{\mathcal{W}}\otimes\mathcal{H}_{\mathcal{Y}}} \lesssim m^2\kappa^4 r_C(\delta,N_1,c_1)^2+\kappa^4\frac{\log(|\mathcal{M}|/\delta)}{N_1^{\frac{1}{2}}N_2}+\kappa^4\frac{\log(|\mathcal{M}|/\delta)}{N_1}.$$

Proof. The proof is a combination of Lemma D.22, Proposition 8 in Mastouri et al. (2021), and a Bernstein inequality. The proof is different from the existing works because $\widehat{p}(y\mid x)$ is involved in this work. See Appendix D.9 for a complete proof.

□

Lemma D.24 (Proposition 11 of Mastouri et al. (2021))**.** *Let* $C_\epsilon=96\ln^2(6/\epsilon)$ *and suppose that* $N_2\geq\frac{2C_\epsilon\mathcal{N}(\lambda_2)}{\lambda_2}$ *and that* $\lambda_2\leq\|T_2\|_{\mathcal{L}(\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t})}$*. Then, w.p.* $1-2\epsilon/3$*,*

$$\begin{aligned}\left\|\widetilde{b}_R^{[t]}\right\|_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} \leq & 4\Bigg(\frac{32\ln^2(6/\epsilon)}{\lambda_2}\left[\frac{\left(c_Y+\left\|b_R^{[t]}\right\|_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}}\right)^2(4+N_2\lambda_2\mathcal{N}(\lambda_2))}{N_2^2\lambda_2}\right] \\ & +\frac{32\ln^2(6/\epsilon)}{\lambda_2}\left[\frac{4\mathcal{B}(\lambda_2)+N_2\mathcal{A}(\lambda_2)}{N_2^2\lambda_2}\right]+\mathcal{B}(\lambda_2)+\left\|b_R^{[t]}\right\|^2_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}}\Bigg)\end{aligned} \tag{114}$$

Following the proof of Proposition 10 in Mastouri et al. (2021), we then have

$$S_{-1} \lesssim \frac{4}{\lambda_2}(m^2\kappa^4 r_C(\delta,N_1,c_1)^2+\kappa^4\frac{\log(|\mathcal{M}|/\delta)}{N_1^{\frac{1}{2}}N_2}+\kappa^4\frac{\log(|\mathcal{M}|/\delta)}{N_1}) \tag{115}$$

and

$$S_0 \leq \frac{4}{\lambda_2}\kappa^6 r_C\left(\delta, N_1, c_1\right)^2 \left\|\widetilde{b}_R^{[t]}\right\|_{\mathcal{H}_{\mathcal{W}}\otimes\mathcal{H}_{\mathcal{Y}}}^2 . \tag{116}$$

Combining all the above results, we then have

$$\begin{aligned}
&S_{-1} = O\left(\frac{r_C\left(\delta, N_1, c_1\right)^2}{\lambda_2} + \frac{1}{\lambda_2 N_1^{\frac{1}{2}} N_2} + \frac{1}{\lambda_2 N_1}\right)\\
&S_0 = O\left(\frac{r_C\left(\delta, N_1, c_1\right)^2}{\lambda_2}\cdot\left(\frac{1}{N_2^2\lambda_2^2} + \frac{1}{N_2\lambda_2^{1+1/b_2}} + \frac{1}{N_2^2\lambda_2^{3-c_2}} + \frac{1}{N_2\lambda_2^{2-c_2}} + \lambda_2^{c_2-1} + 1\right)\right)\\
&\mathcal{A}\left(\lambda_2\right) = O\left(\lambda_2^{c_2}\right),\\
&S_1 = O\left(\frac{1}{N_2^2\lambda_2} + \frac{1}{N_2\lambda_2^{1/b_2}}\right),\\
&S_2 = O\left(\frac{1}{N_2^2\lambda_2^{2-c_2}} + \frac{1}{N_2\lambda_2^{1-c_2}}\right).
\end{aligned} \tag{117}$$

We notice that $\frac{r_C(\delta,N_1,c_1)^2}{\lambda_2}$ dominates $\frac{1}{\lambda_2 N_1}$ in S_{-1}. Furthermore, since $b_2 > 1$ and $c_2 \in (1, 2]$, we have that $\frac{1}{N_2}$ dominates $\frac{1}{N_2\lambda_2^{3-c_2}}$; that $\frac{1}{N_2\lambda_2^{1+1/b_2}}$ dominates $\frac{1}{N_2\lambda_2^{2-c_2}}$; and that 1 dominates $\lambda_2^{c_2-1}$ (since $\lambda_2 \to 0$). In addition, it can be seen that S_1 dominates S_2 for the same reasons.

Therefore, we have

$$\begin{aligned}
&\mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) - \mathcal{L}_R^{[t]}(b_R^{[t]})\\
=&O\left(\frac{r_C\left(\delta, N_1, c_1\right)^2}{\lambda_2}\left[\frac{1}{N_2^2\lambda_2^2} + \frac{1}{N_2\lambda_2^{1+1/b_2}} + 1\right] + \lambda_2^{c_2} + \frac{1}{N_2^2\lambda_2} + \frac{1}{N_2\lambda_2^{1/b_2}} + \frac{1}{\lambda_2 N_1^{\frac{1}{2}} N_2}\right).
\end{aligned} \tag{118}$$

By choosing λ_1, N_1 and λ_2 appropriately, and following the proof of Szabó et al. (2016)[Theorem 5], Mastouri et al. (2021)[Theorem 2], we have the following result:

Fix $\zeta > 0$ and choose $\lambda_1 = N_1^{\frac{1}{c_1+1}}$ and $N_1 = N_2^{\frac{\zeta(c_1+1)}{(c_1-1)}}$.
1. If $\zeta \leq \frac{b_2(c_2+1)}{b_2c_2+1}$, choose $\lambda_2 = N_2^{-\frac{\zeta}{c_2+1}}$. Then $\mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) = O_p\left(N_2^{-\frac{\zeta c_2}{c_2+1}}\right)$.
2. If $\zeta \geq \frac{b_2(c_2+1)}{b_2c_2+1}$, choose $\lambda_2 = N_2^{-\frac{b_2}{b_2c_2+1}}$. Then $\mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) = O_p\left(N_2^{-\frac{b_2c_2}{b_2c_2+1}}\right)$.

In particular, we only consider the optimal rate and the most efficient sample splitting way here. We can let $\zeta = \frac{b_2(c_2+1)}{b_2c_2+1}$, which implies that $\mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) = O_p\left(N_2^{-\frac{b_2c_2}{b_2c_2+1}}\right)$.

Repeating the above arguments T times for $\mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]})$, and $T-1$ times for $\mathcal{L}_D^{[t]}(\widehat{b}_D^{[t]})$, by a union bound argument, it is straightforward to have

$$\max_{t=1:T}\mathcal{L}_R^{[t]}(\widehat{b}_R^{[t]}) = O_P(\log(T)N_2^{-\frac{b_2c_2}{b_2c_2+1}})$$

and

$$\max_{t=1:T-1}\mathcal{L}_D^{[t]}(\widehat{b}_D^{[t]}) = O_P(\log(T)N_2^{-\frac{b_2c_2}{b_2c_2+1}}).$$

The proof is done by defining $\gamma = b_2$.

D.8. Proof of Lemma D.4

We focus on the function $b_{R,t}$ at first, and present the following definition that helps to prove a uniform upper bound.

Definition D.25.

$$\widetilde{\mathcal{L}}_R^{[t]}(b_{R,t}) = \mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b},Y_t\sim\mathrm{Unif}(\mathcal{Y})}\left(\left\langle \widehat{\mu}_{W_t|X_t}\otimes\phi\left(Y_t\right), b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} - \widehat{p}\left(Y_t \mid X_t\right)\right)^2 \tag{119}$$

Recall that we have defined the population risk functional and empirical loss functional in Definition A.2 and (13), which are as follows:

$$\mathcal{L}_R^{[t]}(b_{R,t}) = \mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b},Y_t\sim\mathrm{Unif}(\mathcal{Y})}\left(\left\langle \mu_{W_t|X_t}\otimes\phi\left(Y_t\right), b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} - p\left(Y_t \mid X_t\right)\right)^2 \tag{120}$$

$$\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) = \frac{1}{N_2}\sum_{n=1}^{N_2}\left(\left\langle \widehat{\mu}_{W_t|x'_{t,n}}\otimes\phi\left(y''_{t,n}\right), b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} - \widehat{p}\left(y''_{t,n} \mid x'_{t,n}\right)\right)^2 + \lambda_2\|b_{R,t}\|^2_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}}. \tag{121}$$

Then, we have the following decomposition:

$$\begin{aligned}&\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t})\\ =&\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \widetilde{\mathcal{L}}_R^{[t]}(b_{R,t}) + \widetilde{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t})\end{aligned} \tag{122}$$

and

$$\begin{aligned}&\sup_{b_{R,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}}|\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t})|\\ \leq&\sup_{b_{R,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}}|\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \widetilde{\mathcal{L}}_R^{[t]}(b_{R,t})| + \sup_{b_{R,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}}|\widetilde{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t})|\\ =&I + II\end{aligned} \tag{123}$$

Intuitively, the first term I is related to the stage 2 error, while the second term II is related to the stage 1 error. In order to provide uniform upper bounds on them, the techniques from the subject of empirical process theory can be adopted.

Upper bound term I.

We analyze the term $\sup_{b_{R,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}}|\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \widetilde{\mathcal{L}}_R^{[t]}(b_{R,t})|$ by the empirical process theory. Firstly, we introduce a concept from the empirical process that are used to measure the size of function classes $\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}$. The following definition is adapted from (Wainwright, 2019) and (Foster & Syrgkanis, 2023).

Definition D.26 (Localize population Rademacher complexity and critical radius.)**.** For a real-valued function class $\mathcal{G}$ on a probability space $(\mathcal{X}, P)$, we denote by $\|g\|_2^2$ the expectation of $g(X)^2$, that is $\|g\|_2^2 = \mathbb{E}_{X\sim\mathcal{P}}\left[g(X)^2\right]$. Given any radius $\delta > 0$, the local population Rademacher complexity is given by

$$\mathcal{R}_n(\mathcal{G},\delta) = \mathbb{E}_{\epsilon,X}[\sup_{g\in\mathcal{G}:\|g\|_2\leq\delta}|n^{-1}\sum_{i=1}^n\epsilon_i g\left(X_i\right)|],$$

where $\{X_i\}_{i=1}^n$ are i.i.d. copies of X and $\{\epsilon_i\}_{i=1}^n$ are i.i.d. Rademacher random variables taking values in $\{-1,+1\}$ with equal probability. Further, assume that $\mathcal{G}$ is a 1-uniformly bounded function class $\{g:\mathcal{X}\to\mathbb{R}, \sup_x|g(x)|\leq 1\}$. Further, we assume that $\mathcal{G}$ is a star-shaped function class, i.e. $\alpha g\in\mathcal{G}$ for any $g\in\mathcal{G}$ and scalar $\alpha\in[-1,1]$. Then the critical radius of $\mathcal{G}$, denoted by δ_n, is the solution to the inequality

$$\mathcal{R}_n(\mathcal{G},\delta)\leq\delta^2.$$

In this work, critical radius is used in the theoretical analysis to measure the size of function classes for the bridge functions, which provides a way to get a *uniform* law of large numbers with a convergence rate at each time t.

In particular, we apply Lemma A.5 to upper bound $\sup_{b_{R,t}}|\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \widetilde{\mathcal{L}}_R^{[t]}(b_{R,t})|$. In this case, $\widehat{\mathcal{L}}_R^{[t]}(b_{R,t})$ can be viewed as a regularized empirical loss function for the regularized least square problem. We can use the random variable U_t to

denote $\widehat{p}(Y_t \mid X_t)$ with $X_t \sim \mathbb{P}^{\pi^b}$, $Y_t \sim \text{unif}(\mathcal{Y}_t)$. And the collected i.i.d. samples are $u_{t,k}$, $k = 1, ..., N_2$, which represents $\widehat{p}(y''_{t,n} \mid x_{t,n})$ for $n = 1, ..., N_2$. We note that $\widehat{p}$ only depends on the first sample with sample size N_1, and therefore it does not affect the concentration result at the second stage. Similarly, we can use the randome variable V_t to denote $\widehat{\mu}_{W_t|X_t} \otimes \phi(Y_t)$ with $X_t \sim \mathbb{P}^{\pi^b}$, $Y_t \sim \text{unif}(\mathcal{Y}_t)$. And $V_{t,k}$, $k = 1, ..., N_2$, are i.i.d. samples denoting $\widehat{\mu}_{W_t|x_{t,n}} \otimes \phi(y''_{t,n})$, $n = 1, ..., N_2$. The inner product $\left\langle \widehat{\mu}_{W_t|X_t} \otimes \phi(Y_t), b_{R,t} \right\rangle_{\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}}$ in the Hilbert space $\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}$ can be written as $b_{R,t}[\widehat{\mu}_{W_t|X_t} \otimes \phi(Y_t)] = b_{R,t}[V_t]$. In this perspective, we can view the stage 2 procedure as a regression problem where U_t is the response variable, $b_{R,t}$ is the regression function, and V_t denotes the independent variable. The loss function l is a quadratic loss function with $l(U_t, b[V_t]) = (U_t - b[V_t])^2$.

Then, we have

$$\begin{aligned}
&\sup_{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \widetilde{\mathcal{L}}_R^{[t]}(b_{R,t})| \\
=&\sup_{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathbb{E}}\left[l(U_t, b_{R,t}[V_t])\right] + \lambda_2 \|b_{R,t}\|_{\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} - \mathbb{E}\left[l(U_t, b_{R,t}[V_t])\right]| \\
\leq&\sup_{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathbb{E}}\left[l(U_t, b_{R,t}[V_t])\right] - \mathbb{E}\left[l(U_t, b_{R,t}[V_t])\right]| + \lambda_2 \sup_{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} \|b_{R,t}\|_{\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}}
\end{aligned} \tag{124}$$

We then apply Lemma A.5 to provide $\sup_{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathbb{E}}\left[l(U_t, b_{R,t}[V_t])\right] - \mathbb{E}\left[l(U_t, b_{R,t}[V_t])\right]|$ an upper bound. In particular, we let δ_{R,N_2} be the critical radius (See Definition D.26) of the function class $\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}$ for $b_{R,t}$, depending on the stage 2 sample size N_2. To see the function $l(U_t, b_{R,t}[V_t])$ is Lipschitz continuous with respect to $b_{R,t}$, we do a direct calculation in the following.

$$\begin{aligned}
&|l(U_t, b_{R,t}[V_t]) - l(U_t, b'_{R,t}[V_t])| \\
=&|(U_t, b_{R,t}[V_t])^2 - (U_t, b'_{R,t}[V_t])^2| \\
=&|b_{R,t}[V_t]^2 - b'_{R,t}[V_t]^2 + 2U_t b_{R,t}[V_t] - 2U_t, b'_{R,t}[V_t]| \\
\leq&|b_{R,t}[V_t]^2 - b'_{R,t}[V_t]^2| + 2|U_t||b_{R,t}[V_t] - b'_{R,t}[V_t]| \\
=&|b_{R,t}[V_t] + b'_{R,t}[V_t]||b_{R,t}[V_t] - b'_{R,t}[V_t]| + 2|U_t||b_{R,t}[V_t] - b'_{R,t}[V_t]| \\
\leq&(|b_{R,t}[V_t]| + |b'_{R,t}[V_t]| + 2|U_t|)|b_{R,t}[V_t] - b'_{R,t}[V_t]|
\end{aligned} \tag{125}$$

By Assumption D.16, we have $|b_{R,t}[V_t]| = |\left\langle \mu_{W_t|X_t} \otimes \phi(Y_t), b_{R,t} \right\rangle_{\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}}| \leq \|\mu_{W_t|X_t}\|_{\mathcal{H}_{\mathcal{W}_t}} \|\phi(Y_t)\|_{\mathcal{H}_{\mathcal{Y}_t}} \|b_{R,t}\|_{\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} \leq \kappa^2$. In addition, by Assumption D.16, we have $|U_t| \leq m$. Therefore, l is Lipschitz continuous with respect to the function $b_{R,t}$ with a Lipschitz constant $2\kappa^2 + 2m$.

By applying Lemma A.5 and that $\sup_{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} \|b_{R,t}\|_{\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} \leq 1$, with probability at least $1 - c_5 \exp\left(c_6 N_2 \delta_{R,N_2}^2\right)$, we have

$$\begin{aligned}
&\sup_{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \widetilde{\mathcal{L}}_R^{[t]}(b_{R,t})| \\
\leq&\sup_{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathbb{E}}\left[l(U_t, b_{R,t}[V_t])\right] - \mathbb{E}\left[l(U_t, b_{R,t}[V_t])\right]| + \lambda_2 \sup_{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} \|b_{R,t}\|_{\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} \\
\leq&36(\kappa^2 + m)\delta_{R,N_2} \left(\sup_{b_{R,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}} \|b_{R,t}\|_2 + \delta_{R,N_2} \right) + \lambda_2 \\
\leq&36(\kappa^2 + m)\delta_{R,N_2} \left(M_R + \delta_{R,N_2}\right) + \lambda_2
\end{aligned} \tag{126}$$

where the final inequality is from Assumption D.16 with M_R being the uniform upper bound on the function space $\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}$ under the $\|\cdot\|_\infty$ perspective.

The following lemma provides an upper bound on the critical radius δ_{R,N_2} of the RKHS $\mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Y}_t}$.

Lemma D.27 (Corollary 14.5 of Wainwright (2019)). *Let $\mathcal{F} = \{f \in \mathbb{H} \mid \|f\|_H \leq 1\}$ be the unit ball of an RKHS with*

eigenvalues $(\mu_j)_{j=1}^{\infty}$. *Then the localized population Rademacher complexity (see Definition D.26) is upper bounded as*

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

In this work, we are focusing on the eigenvalues of the covariance operator T_2, which are denoted as $(l_k)_{k\in\mathbb{N}}$. The eigenvalues of T_2 can also be viewed as the eigenvalues of our kernel in the considered RKHS. See more discussions on this relationship in Definition 1 and Remark 2 of Caponnetto & De Vito (2007).

According to Assumption D.16, the eigenvalues decay as of the order $l_k \sim k^{-b_2}$ for some $b_2 > 1$. Such decay-rate is usually named as polynomial eigen-decay rate. We note that the polynomial eigen-decay rate for RKHSs is commonly considered in practice (e.g. $b_2/2$-order Sobolev space). In particular, larger b_2 means faster decay of the eigenvalues of the covariance operator T_2, and smaller effective input dimension.

Following the calculation in Example 13.20 of Wainwright (2019), it can be seen that

$$\delta_{R,N_2} \asymp N_2^{-\frac{b_2}{2b_2+2}}.$$

Consequently, with probability at least $1 - c_5\exp\left(c_6 N_2 \delta_{R,N_2}^2\right)$, we have

$$\begin{aligned}
&\sup_{b_{R,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \widetilde{\mathcal{L}}_R^{[t]}(b_{R,t})| \\
\leq& 36(\kappa^2+m)\delta_{R,N_2}\left(M_R + \delta_{R,N_2}\right) + \lambda_2 \\
\lesssim& (\kappa^2+m)M_R N_2^{-\frac{b_2}{2b_2+2}} + \lambda_2.
\end{aligned} \tag{127}$$

Upper bound term II.

Next, we analyze the term $\widetilde{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t})$. Assumptions in Lemma D.5 can be applied here.

By definition, we have

$$\begin{aligned}
&\widetilde{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t}) \\
=&\mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b},Y_t\sim\mathrm{Unif}(\mathcal{Y})}\left(\left\langle\widehat{\mu}_{W_t|X_t}\otimes\phi\left(Y_t\right),b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} - \widehat{p}\left(Y_t\mid X_t\right)\right)^2 \\
&- \mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b},Y_t\sim\mathrm{Unif}(\mathcal{Y})}\left(\left\langle\mu_{W_t|X_t}\otimes\phi\left(Y_t\right),b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} - p\left(Y_t\mid X_t\right)\right)^2 \\
=&\mathbb{E}\left(\left\langle\widehat{\mu}_{W_t|X_t}\otimes\phi\left(Y_t\right),b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} - \widehat{p}\left(Y_t\mid X_t\right) + \left\langle\mu_{W_t|X_t}\otimes\phi\left(Y_t\right),b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} - p\left(Y_t\mid X_t\right)\right) \\
&\left(\left\langle\widehat{\mu}_{W_t|X_t}\otimes\phi\left(Y_t\right),b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} - \widehat{p}\left(Y_t\mid X_t\right) - \left\langle\mu_{W_t|X_t}\otimes\phi\left(Y_t\right),b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} + p\left(Y_t\mid X_t\right)\right) \\
:=& \mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b},Y_t\sim\mathrm{Unif}(\mathcal{Y})}[A(X_t,Y_t)B(X_t,Y_t)]
\end{aligned} \tag{128}$$

where we use the notations as follows:

$$A(X_t,Y_t) = \left\langle\widehat{\mu}_{W_t|X_t}\otimes\phi\left(Y_t\right),b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} - \widehat{p}\left(Y_t\mid X_t\right) + \left\langle\mu_{W_t|X_t}\otimes\phi\left(Y_t\right),b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} - p\left(Y_t\mid X_t\right) \tag{129}$$

$$B(X_t,Y_t) = \left\langle\widehat{\mu}_{W_t|X_t}\otimes\phi\left(Y_t\right),b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} - \widehat{p}\left(Y_t\mid X_t\right) - \left\langle\mu_{W_t|X_t}\otimes\phi\left(Y_t\right),b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} + p\left(Y_t\mid X_t\right). \tag{130}$$

We then have

$$\mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b},Y_t\sim\mathrm{Unif}(\mathcal{Y})}[A(X_t,Y_t)B(X_t,Y_t)] \leq \mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b},Y_t\sim\mathrm{Unif}(\mathcal{Y})}[|A(X_t,Y_t)||B(X_t,Y_t)|]. \tag{131}$$

For $|A(X_t,Y_t)|$, we notice that $|A(X_t,Y_t)| \leq |\left\langle\widehat{\mu}_{W_t|X_t}\otimes\phi\left(Y_t\right),b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}}| + |\left\langle\mu_{W_t|X_t}\otimes\phi\left(Y_t\right),b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}}| + 2m$ by Assumption D.16. In addition, we have

$$|\left\langle\widehat{\mu}_{W_t|X_t}\otimes\phi\left(Y_t\right),b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}}| \leq \|\widehat{\mu}_{W_t|X_t}\|_{\mathcal{H}_{\mathcal{W}_t}}\|\phi\left(Y_t\right)\|_{\mathcal{H}_{\mathcal{Y}_t}}\|b_{R,t}\|_{\mathcal{H}_{\mathcal{W}_t}\times\mathcal{H}_{\mathcal{Y}_t}} \leq \kappa^2$$

and similarly $|\left\langle \mu_{W_t|X_t} \otimes \phi\left(Y_t\right), b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}}| \leq \kappa^2$. Therefore, we have $|A(X_t, Y_t)| \leq 2\kappa^2 + 2m$. Consequently, the following inequality holds.

$$\mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b}, Y_t\sim\mathrm{Unif}(\mathcal{Y})}[A(X_t,Y_t)B(X_t,Y_t)] \leq (2\kappa^2+2m)\mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b}, Y_t\sim\mathrm{Unif}(\mathcal{Y})}[|B(X_t,Y_t)|].$$

Furthermore, with probability at least $1-2\delta$, we have

$$\begin{aligned}
&\mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b}, Y_t\sim\mathrm{Unif}(\mathcal{Y})}[|B(X_t,Y_t)|]\\
=&\mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b}, Y_t\sim\mathrm{Unif}(\mathcal{Y})}[|\left\langle \widehat{\mu}_{W_t|X_t} \otimes \phi\left(Y_t\right), b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} -\\
&\widehat{p}\left(Y_t \mid X_t\right) - \left\langle \mu_{W_t|X_t} \otimes \phi\left(Y_t\right), b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} + p\left(Y_t \mid X_t\right)|]\\
\leq&\mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b}, Y_t\sim\mathrm{Unif}(\mathcal{Y})}[|\left\langle \mu_{W_t|X_t} \otimes \phi\left(Y_t\right), b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} - \left\langle \widehat{\mu}_{W_t|X_t} \otimes \phi\left(Y_t\right), b_{R,t}\right\rangle_{\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}}|]\\
&+\mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b}, Y_t\sim\mathrm{Unif}(\mathcal{Y})}[|\widehat{p}(Y_t \mid X_t) - p(Y_t \mid X_t)|]\\
\leq&\kappa\mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b}}\|\widehat{\mu}_{W_t|X_t} - \mu_{W_t|X_t}\|_{\mathcal{H}_{\mathcal{W}_t}} + \frac{1}{\mathrm{vol}(\mathcal{Y}_t)}\mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b}}\mathrm{TV}(\widehat{p}(\cdot \mid X_t), p(\cdot \mid X_t))\\
\lesssim&\kappa^2 r_C(\delta, N_1, c_1) + \frac{1}{\mathrm{vol}(\mathcal{Y}_t)}\mathbb{E}_{X_t\sim\mathbb{P}^{\pi^b}}\mathrm{TV}(\widehat{p}(\cdot \mid X_t), p(\cdot \mid X_t))\\
\lesssim&\kappa^2\frac{\sqrt{\gamma_1}\left(c_1+1\right)}{4^{\frac{1}{c_1+1}}}\left(\frac{4\kappa\left(\kappa+\kappa\left\|C_{W|X}\right\|_{\mathcal{H}_\Gamma}\right)\ln(2/\delta)}{\sqrt{N_1\gamma_1}\left(c_1-1\right)}\right)^{\frac{c_1-1}{c_1+1}} + \frac{\ln(|\mathcal{M}|/\delta)}{\sqrt{N_1}}
\end{aligned} \tag{132}$$

where the last two inequalities come from the estimation error from stage 1 (Theorem D.12 and Lemma D.22).

Therefore, with probability at least $1-2\delta$, we have

$$\begin{aligned}
&\sup_{b_{R,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} |\widetilde{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t})|\\
\lesssim&(2\kappa^2+2m)\kappa^2\frac{\sqrt{\gamma_1}\left(c_1+1\right)}{4^{\frac{1}{c_1+1}}}\left(\frac{4\kappa\left(\kappa+\kappa\left\|C_{W|X}\right\|_{\mathcal{H}_\Gamma}\right)\ln(2/\delta)}{\sqrt{N_1\gamma_1}\left(c_1-1\right)}\right)^{\frac{c_1-1}{c_1+1}} + (2\kappa^2+2m)\frac{\ln(|\mathcal{M}|/\delta)}{\sqrt{N_1}}.
\end{aligned} \tag{133}$$

Combining I and II.

By combining the upper bounds on I and II, with probability at least $1-2\delta-c_5\exp\left(c_6 N_2\delta_{R,N_2}^2\right)$, we have

$$\begin{aligned}
&\sup_{b_{R,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t})|\\
\leq&I+II\\
\lesssim&(\kappa^2+m)M_R N_2^{-\frac{b_2}{2b_2+2}}\\
&+(\kappa^2+m)\kappa^2\frac{\sqrt{\gamma_1}\left(c_1+1\right)}{4^{\frac{1}{c_1+1}}}\left(\frac{4\kappa\left(\kappa+\kappa\left\|C_{W|X}\right\|_{\mathcal{H}_\Gamma}\right)\ln(2/\delta)}{\sqrt{N_1\gamma_1}\left(c_1-1\right)}\right)^{\frac{c_1-1}{c_1+1}} + (\kappa^2+m)\frac{\ln(|\mathcal{M}|/\delta)}{\sqrt{N_1}}.
\end{aligned} \tag{134}$$

According to a direct computation with the sample splitting procedure $N_1 = N_2^{\frac{c_1+1}{c_1-1}\frac{\gamma(c_2+1)}{\gamma c_2+1}}$, we note that the first term is the dominating term, which is of the order $N_2^{-\frac{\gamma}{2\gamma+2}}$ by setting γ as b_2. Furthermore, by repeating the above argument T times for $\sup_{b_{R,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t})|$ and $T-1$ times for $\sup_{b_{D,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Z}_t}} |\widehat{\mathcal{L}}_D^{[t]}(b_{D,t}) - \mathcal{L}_D^{[t]}(b_{D,t})|$, it can be seen that

$$\max_{t=1:T}\sup_{b_{R,t}\in\mathcal{H}_{\mathcal{W}_t}\otimes\mathcal{H}_{\mathcal{Y}_t}} |\widehat{\mathcal{L}}_R^{[t]}(b_{R,t}) - \mathcal{L}_R^{[t]}(b_{R,t})| = O_P(M_R\log(T/\delta)N_2^{-\frac{\gamma}{2\gamma+2}})$$

and

$$\max_{t=1:T-1} \sup_{b_{D,t} \in \mathcal{H}_{\mathcal{W}_t} \otimes \mathcal{H}_{\mathcal{Z}_t}} |\widehat{\mathcal{L}}_D^{[t]}(b_{D,t}) - \mathcal{L}_D^{[t]}(b_{D,t})| = O_P(M_D \log(T/\delta) N_2^{-\frac{\gamma}{2\gamma+2}}).$$

The proof is done.

D.9. Proof of Lemma D.23

Proof. We have

$$\begin{aligned}
&\|\widehat{g}_2 - g_2\|_{\mathcal{H}_\mathcal{W} \otimes \mathcal{H}_\mathcal{Y}} \\
=&\left\| \frac{1}{N_2} \sum_{n=1}^{N_2} \left[\widehat{\mu}_{W|x'_n} \otimes \phi\left(y''_{t,n}\right)\right] \widehat{p}(y''_{t,n} \mid x'_n) - \frac{1}{N_2} \sum_{n=1}^{N_2} \left[\mu_{W|x'_n} \otimes \phi\left(y''_{t,n}\right)\right] p(y''_{t,n} \mid x'_n) \right\|_{\mathcal{H}_\mathcal{W} \otimes \mathcal{H}_\mathcal{Y}} \\
\leq&\left\| \frac{1}{N_2} \sum_{n=1}^{N_2} \left[\widehat{\mu}_{W|x'_n} \otimes \phi\left(y''_{t,n}\right)\right] \widehat{p}(y''_{t,n} \mid x'_n) - \frac{1}{N_2} \sum_{n=1}^{N_2} \left[\widehat{\mu}_{W|x'_n} \otimes \phi\left(y''_{t,n}\right)\right] p(y''_{t,n} \mid x'_n) \right\|_{\mathcal{H}_\mathcal{W} \otimes \mathcal{H}_\mathcal{Y}} \\
&+ \left\| \frac{1}{N_2} \sum_{n=1}^{N_2} \left[\widehat{\mu}_{W|x'_n} \otimes \phi\left(y''_{t,n}\right)\right] p(y''_{t,n} \mid x'_n) - \frac{1}{N_2} \sum_{n=1}^{N_2} \left[\mu_{W|x'_n} \otimes \phi\left(y''_{t,n}\right)\right] p(y''_{t,n} \mid x'_n) \right\|_{\mathcal{H}_\mathcal{W} \otimes \mathcal{H}_\mathcal{Y}} \\
=&I + II
\end{aligned} \tag{135}$$

Then,

$$\begin{aligned}
&I \\
=&\left\| \frac{1}{N_2} \sum_{n=1}^{N_2} \left[\widehat{\mu}_{W|x'_n} \otimes \phi\left(y''_{t,n}\right)\right] \widehat{p}(y''_{t,n} \mid x'_n) - \frac{1}{N_2} \sum_{n=1}^{N_2} \left[\widehat{\mu}_{W|x'_n} \otimes \phi\left(y''_{t,n}\right)\right] p(y''_{t,n} \mid x'_n) \right\|_{\mathcal{H}_\mathcal{W} \otimes \mathcal{H}_\mathcal{Y}} \\
\leq&\frac{1}{N_2} \sum_{n=1}^{N_2} \left\| \left[\widehat{\mu}_{W|x'_n} \otimes \phi\left(y''_{t,n}\right)\right] \left(\widehat{p}(y''_{t,n} \mid x'_n) - p(y''_{t,n} \mid x'_n)\right) \right\|_{\mathcal{H}_\mathcal{W} \otimes \mathcal{H}_\mathcal{Y}} \\
\leq&\frac{1}{N_2} \sum_{n=1}^{N_2} \left\| \left[\widehat{\mu}_{W|x'_n}\right] \right\|_{\mathcal{H}_\mathcal{W}} \|\phi\left(y''_{t,n}\right)\|_{\mathcal{H}_\mathcal{Y}} |\widehat{p}(y''_{t,n} \mid x'_n) - p(y''_{t,n} \mid x'_n)| \\
\leq&\kappa^2 \frac{1}{N_2} \sum_{n=1}^{N_2} |\widehat{p}(y''_{t,n} \mid x'_n) - p(y''_{t,n} \mid x'_n)| \\
=&\kappa^2 \left(\frac{1}{N_2} \sum_{n=1}^{N_2} |\widehat{p}(y''_{t,n} \mid x'_n) - p(y''_{t,n} \mid x'_n)| - \frac{1}{\mathrm{vol}(\mathcal{Y})} \mathbb{E}_{x \sim \mathbb{P}^{\pi^b}} \mathrm{TV}\left(\widehat{P}_{\mathrm{MLE}}(\cdot \mid x), P^\star(\cdot \mid x)\right) \right) \\
&+ \kappa^2 \frac{1}{\mathrm{vol}(\mathcal{Y})} \mathbb{E}_{x \sim \mathbb{P}^{\pi^b}} \mathrm{TV}\left(\widehat{P}_{\mathrm{MLE}}(\cdot \mid x), P^\star(\cdot \mid x)\right) \\
=&\kappa^2 \left(\frac{1}{N_2} \sum_{n=1}^{N_2} |\widehat{p}(y''_{t,n} \mid x'_n) - p(y''_{t,n} \mid x'_n)| - \mathbb{E}_{X \sim \mathbb{P}^{\pi^b}, Y \sim \mathrm{unif}(\mathcal{Y})} |\widehat{p}(Y \mid X) - p(Y \mid X)| \right) \\
&+ \kappa^2 \frac{1}{\mathrm{vol}(\mathcal{Y})} \mathbb{E}_{x \sim \mathbb{P}^{\pi^b}} \mathrm{TV}\left(\widehat{P}_{\mathrm{MLE}}(\cdot \mid x), P^\star(\cdot \mid x)\right) \\
\lesssim&\kappa^2 \sqrt{\frac{\log(|\mathcal{M}|/\delta)}{N_1^{\frac{1}{2}} N_2}} + \kappa^2 \sqrt{\frac{\log(|\mathcal{M}|/\delta)}{N_1}} \text{ with probability at least } 1 - 2\delta,
\end{aligned} \tag{136}$$

where we apply a Bernstein inequality for the first term, and the convergence rate of the standard MLE (Lemma D.22) for the second term.

$$
\begin{aligned}
&II \\
=&\left\| \frac{1}{N_2} \sum_{n=1}^{N_2} \left[\widehat{\mu}_{W|x'_n} \otimes \phi\left(y''_{t,n}\right) \right] p(y''_{t,n} \mid x'_n) - \frac{1}{N_2} \sum_{n=1}^{N_2} \left[\mu_{W|x'_n} \otimes \phi\left(y''_{t,n}\right) \right] p(y''_{t,n} \mid x'_n) \right\|_{\mathcal{H}_{\mathcal{W}} \otimes \mathcal{H}_{\mathcal{Y}}} \\
\leq& \frac{1}{N_2} \sum_{n=1}^{N_2} \left\| \left[(\widehat{\mu}_{W|x'_n} - \mu_{W|x'_n}) \otimes \phi\left(y''_{t,n}\right) \right] p(y''_{t,n} \mid x'_n) \right\|_{\mathcal{H}_{\mathcal{W}} \otimes \mathcal{H}_{\mathcal{Y}}} \\
\leq& \frac{1}{N_2} \sum_{n=1}^{N_2} \left\| \left[(\widehat{\mu}_{W|x'_n} - \mu_{W|x'_n}) \right] \right\|_{\mathcal{H}_{\mathcal{W}}} \left\| \left[\phi\left(y''_{t,n}\right) \right] \right\|_{\mathcal{H}_{\mathcal{Y}}} |p(y''_{t,n} \mid x'_n)| \\
\leq& m\kappa^2 r_C\left(\delta, N_1, c_1\right) \text{ with probability at least } 1-\delta,
\end{aligned}
\tag{137}
$$

where we apply the finite sample rate for stage 1 estimation (Lemma D.12) and Assumption D.16 in the final inequality. Then, we have $\|\widehat{g}_2 - g_2\|^2_{\mathcal{H}_{\mathcal{W}} \otimes \mathcal{H}_{\mathcal{Y}}} \leq 2(I)^2 + 2(II)^2 \lesssim m^2\kappa^4 r_C\left(\delta, N_1, c_1\right)^2 + \kappa^4 \frac{\log(|\mathcal{M}|/\delta)}{N_1^{\frac{1}{2}} N_2} + \kappa^4 \frac{\log(|\mathcal{M}|/\delta)}{N_1}$ with probability at least $1 - 3\delta$. □