Off-Policy Evaluation for Episodic Partially Observable Markov Decision Processes under Non-Parametric Models

Rui Miao
University of California, Irvine
rmiao2@uci.edu

Zhengling Qi*
The George Washington University
qizhengling@gwu.edu

Xiaoke Zhang
The George Washington University
xkzhang@gwu.edu

Abstract

We study the problem of off-policy evaluation (OPE) for episodic Partially Observable Markov Decision Processes (POMDPs) with continuous states. Motivated by the recently proposed proximal causal inference framework, we develop a non-parametric identification result for estimating the policy value via a sequence of so-called *V-bridge* functions with the help of time-dependent proxy variables. We then develop a fitted-Q-evaluation-type algorithm to estimate V-bridge functions recursively, where a non-parametric instrumental variable (NPIV) problem is solved at each step. By analyzing this challenging sequential NPIV problem, we establish the finite-sample error bounds for estimating the V-bridge functions and accordingly that for evaluating the policy value, in terms of the sample size, length of horizon and so-called *(local) measure of ill-posedness* at each step. To the best of our knowledge, this is the first finite-sample error bound for OPE in POMDPs under non-parametric models.

1 Introduction

In practical reinforcement learning (RL), a representation of the full state which makes the system Markovian and therefore amenable to most existing RL algorithms is not known *a priori.* Decision makers are often facing so-called *partial observability* of the state information, which significantly hinders the task of RL. In general, agents have to maintain all historical information and establish a belief system on the hidden state for optimal decision making. A partially observable Markov decision process (POMDP) is often used to model the data generating process. See examples in robotics [Rafferty et al., 2011], precision medicine [Tsoukalas et al., 2015], stochastic game [Hansen et al., 2004] and many others. However, it is well known that learning optimal policies in POMDP is computationally intractable [Papadimitriou and Tsitsiklis, 1987]. The issue of partial observability becomes more serious in the batch setting, where agents are not able to actively collect additional data and further explore the environment. For example, standard off-policy evaluation (OPE) methods, which aim to learn a policy value from the batch data generated from some behavior policy, would fail to give a consistent estimate because of unobserved state variables.

Due to this practical concern, there is a recent line of research studying the OPE under the framework of a confounded POMDP, where the behavior policy to generate the batch data is allowed to depend

*Corresponding author.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

https://doi.org/10.52202/068431-0043

on some unobserved state variables [e.g., Tennenholtz et al., 2020, Nair and Jiang, 2021, Bennett and Kallus, 2021, Shi et al., 2021]. Their identification results on the policy value are inspired by the negative controls or so-called proxy variables in the literature of causal inference [e.g., Miao et al., 2018a, Tchetgen Tchetgen et al., 2020]. A building block of these results is the existence of some bridge functions, namely Q/V-bridge or weight-bridge functions, which are projections of the Q/V-functions or importance weights defined over the original state space onto the observation space. The corresponding statistical estimation of these bridge functions mainly relies on solving linear integral equations [e.g., Kress, 1989]. Different from the tabular case studied by Tennenholtz et al. [2020] and Nair and Jiang [2021] and linear models studied by Shi et al. [2021] theoretically, solving linear integral equations with non-parametric models in the continuous state/observation space are known to be challenging due to the potential ill-posedness [Chen and Reiss, 2011], leading to slow statistical convergence rates. However, existing theoretical results developed by Bennett and Kallus [2021] and Shi et al. [2021] require fast enough convergence rates for these bridge function estimators in order to establish the asymptotic normality of their estimators for OPE, which could be illusive when the problem is seriously ill-posed under non-parametric models. This is different from the supervised learning where a fast enough convergence rate can be easily achieved under non-parametric models. Therefore, to fill this important theoretical gap, it is necessary to study the finite-sample performance of OPE of which bridge functions are estimated non-parametrically.

Motivated by these, in this paper, we study the OPE for confounded and episodic POMDPs with continuous states, where we non-parametrically estimate V-bridge functions. Our main contribution to the literature is three-fold. First, relying on some time-dependent proxy variables, we establish a non-parametric identification result for OPE using V-bridge functions for time-inhomogeneous confounded POMDPs. Based on the identification result, we develop a new fitted-Q-evaluation(FQE)-type approach to estimating V-bridge functions recursively and obtain an estimator for OPE based on the bridge function estimators. At each step of our algorithm, we propose to fit a non-parametric instrumental variable (NPIV) regression using a min-max estimation method, i.e., solving a linear integral equation with a non-parametric model. Our algorithm can be viewed as a sequential NPIV estimation, which is not well studied in the literature. Second and most importantly, we establish the finite-sample error bound for estimating V-bridge functions and accordingly that for evaluating the policy value, in terms of the *sample size*, *length of horizon* and *(local) measure of ill-posedness* at each step. Unlike the well studied standard NPIV model in the econometrics literature [e.g., Ai and Chen, 2003, Newey and Powell, 2003] where the response variable is directly observed, the response variable in our NPIV model at each step of the algorithm relies on the model estimate at its previous step. This difference makes our theoretical analysis substantially difficult. By carefully characterizing the statistical error due to the NPIV estimation at each step and more importantly, its propagation effect on future estimates, we are able to establish the first finite-sample result of OPE for confounded POMDPs under non-parametric models, which achieves a polynomial order over the length of horizon and sample size. Finally, our theoretical results on the sequential NPIV estimation are generally applicable to other sequential-type conditional moment restriction problems. The development of the uniform finite-sample error bounds of the NPIV estimation, extending the pointwise result in the previous literature such as Dikkala et al. [2020], may be of independent interest.

2 Related Work

Recently there is a surge of interest in studying OPE with unobserved variables in the sequential decision making problem. Specifically, Zhang and Bareinboim [2016] are among the first who proposed the framework of confounded MDPs, which essentially considers i.i.d. confounders in the dynamic system and therefore preserves the Markovian property. Along this direction, OPE methods are developed under various identification conditions such as partial identification using sensitivity analysis [Namkoong et al., 2020, Kallus and Zhou, 2020, Bruns-Smith, 2021], instrumental variable or mediator assisted OPE [Liao et al., 2021, Li et al., 2021, Shi et al., 2022] and many others. Another line of research focuses on more general confounded POMDP models , where the Markovian assumption is violated, under which several point estimation results were developed such as the aforementioned proxy variables related methods [Tennenholtz et al., 2020, Deaner, 2018, Ying et al., 2021, Bennett and Kallus, 2021, Nair and Jiang, 2021, Shi et al., 2021], spectral methods in undercomplete POMDPs [Hsu et al., 2012, Anandkumar et al., 2014, Jin et al., 2020] and predictive state representation related methods [Littman and Sutton, 2001, Singh et al., 2012, Cai et al., 2022].

https://doi.org/10.52202/068431-0043

Our proposed method, which uses proxy variables for OPE, is closely related to those recently developed by Bennett and Kallus [2021], Shi et al. [2021], and Ying et al. [2021]. Bennett and Kallus [2021] and Ying et al. [2021] studied episodic POMDPs (or complex longitudinal studies) and mainly focused on developing asymptotic normality results of their policy value estimators. Their results rely on some high level rate conditions on the bridge function estimation, which are *unknown* if they would be satisfied when using non-parametric models due to the aforementioned measure of ill-posedness. Shi et al. [2021] mainly focused on time-homogeneous infinite-horizon POMDPs and developed asymptotic normality for their estimators under similar high-level conditions, which therefore has the same issue. Besides, while Shi et al. [2021] also established finite-sample bounds for their bridge function estimation and corresponding OPE, they only study the tabular case or linear/parametric models, where the issue of ill-posedness *does not exist*. In this paper, we provide a systematic investigation on the estimation of V-bridge functions and establish finite-sample guarantees for them and the corresponding OPE under non-parametric models. Specifically, we tackle the challenging episodic setting, where V-bridge functions are estimated sequentially. Without carefully controlling the effect of ill-posedness at each step and its propagation effect on future steps, the estimation error for these V-bridge functions and also that for OPE could be exponentially large in terms of the length of horizon. Motivated by the chaining argument in the empirical process theory, we successfully disentangle the effects of ill-posedness on the current step and future steps separately and thus establish finite-sample bounds for V-bridge functions and OPE both with a polynomial dependence on the length of horizon, which are new theoretical results we contribute to the literature.

Since our V-bridge function estimation can be formulated as a sequential NPIV problem, it is natually related to classical NPIV estimations, which have been extensively studied in the econometrics literature [see, e.g., Newey and Powell, 2003, Ai and Chen, 2003, 2012, Hall and Horowitz, 2005, Chen and Reiss, 2011, Chen and Christensen, 2018, Darolles et al., 2011, Blundell et al., 2007, for earlier reference]. Recently there is also a growing interest in the min-max estimation for NPIV models [see, e.g., Muandet et al., 2020, Dikkala et al., 2020, Hartford et al., 2017, for some recent developments]. As commented before, existing theoretical results for standard NPIV models cannot be directly applied to our setting due to the sequential structure of our FQE algorithm, so we need to develop new theory to address our setting. Technically, in order to establish a polynomial-order finite-sample error bound over the length of horizon for OPE, which is particularly important in RL, we decompose the measure of ill-posedness at each step of our sequential NPIV estimation into two components: the so-called (local) measure of one-step transition ill-posedness and the standard (local) measure of ill-posedness [e.g., Chen and Pouzo, 2012]. Thanks to this novel decomposition, the effect of the first component on the estimation error of V-bridge functions and OPE is multiplicative but can be properly controlled while that of the second component could be large but is only cumulative. See Theorem 6.1. Finally, we remark that Ai and Chen [2012] also studied the sequential NPIV estimation problem, where the non-parametric components are estimated jointly. However, this method could be computationally inefficient in RL with a long horizon. More importantly, their results are built on the nested structure among conditional moment restriction models, which are not satisfied in our setting.

3 Preliminaries and Notations

In this section, we introduce the framework of discrete-time confounded POMDPs and its related OPE problem. Consider an episodic and confounded POMDP denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{U}, \mathcal{A}, T, \mathcal{P}, r)$, with $\mathcal{S}$ and $\mathcal{U}$ as the observed and unobserved continuous state spaces respectively, $\mathcal{A}$ as the discrete action space, T as the length of horizon, $\mathcal{P} = \{\mathbb{P}_t\}_{t=1}^T$ as the transition kernel over $\mathcal{S} \times \mathcal{U} \times \mathcal{A}$ to $\mathcal{S} \times \mathcal{U}$, and $r = \{r_t\}_{t=1}^T$ as the reward function over $\mathcal{S} \times \mathcal{U} \times \mathcal{A}$. $\mathcal{S}$ can also be treated as the observation space in the classical POMDP. Then the process of $\mathcal{M}$ can be summarized as $\{S_t, U_t, A_t, R_t\}_{t=1}^T$ with S_t and U_t as observed and unobserved state variables, A_t as the action, and R_t as the reward, where $r_t(s, u, a) = \mathbb{E}[R_t \mid S_t = s, U_t = u, A_t = a]$ for any $(s, u, a) \in \mathcal{S} \times \mathcal{U} \times \mathcal{A}$. For simplicity, we assume that $|R_t| \leq 1$ uniformly in $1 \leq t \leq T$.

The goal of OPE in a confounded POMDP is to evaluate the performance of a target policy using the batch data collected by some behavior policy. In this paper, the target policy we focus on is a sequence of functions mapping from the state space $\mathcal{S}$ to a probability mass function over the action space $\mathcal{A}$, denoted by $\pi = \{\pi_t\}_{t=1}^T$, where $\pi_t(a \mid s)$ is the probability of choosing an action $A_t = a$ given the state value $S_t = s$. We remark that our proposed identification results stated in Section 4 can be generalized to other policies such as history-dependent ones. Given a target policy π, define

https://doi.org/10.52202/068431-0043

its state value function as

$$V_t^\pi(s,u) = \mathbb{E}^\pi[\textstyle\sum_{t'=t}^T R_{t'} \mid S_t = s, U_t = u], \quad \text{for every } (s,u) \in \mathcal{S} \times \mathcal{U}, \tag{1}$$

where $\mathbb{E}^\pi$ denotes the expectation with respect to the distribution whose action at decision time t follows π_t for any $t \geq 1$. We consider the batch setting, where the observed action A_t is generated by some behavior policy $\tilde{\pi}_t^b$ depending on both S_t and U_t for $1 \leq t \leq T$. We aim to use the batch data to estimate the *policy value* of a target policy π, which is defined as

$$\mathcal{V}(\pi) = \mathbb{E}[V_1^\pi(S_1, U_1)], \tag{2}$$

where $\mathbb{E}$ denotes the expectation with respect to the behavior policy. Due to the unobserved U_t, standard OPE methods that rely on the Markovianity will give bias estimations. In the following, we introduce an identification result for estimating the policy value using some proxy variables.

Notations: For two sequences $\{\varpi(n)\}_{n\geq 1}$ and $\{\theta(n)\}_{n\geq 1}$, the notation $\varpi(n) \gtrsim \theta(n)$ (resp. $\varpi(n) \lesssim \theta(n)$) means that there exists a sufficiently large constant (resp. small) constant $c_1 > 0$ (resp. $c_2 > 0$) such that $\varpi(N) \geq c_1\theta(N)$ (resp. $\varpi(n) \leq c_2\theta(n)$). We use $\varpi(n) \asymp \theta(n)$ when $\varpi(n) \gtrsim \theta(n)$ and $\varpi(n) \lesssim \theta(n)$. For any random variable X, we use $\mathcal{L}^q\{X\}$ to denote the class of all measurable functions with finite q-th moments for $1 \leq q \leq \infty$. Then the $\mathcal{L}^q$-norm is denoted by $\| \bullet \|_{\mathcal{L}^q\{X\}}$. When there is no confusion in the underlying distribution, we also write it as $\| \bullet \|_{\mathcal{L}^q}$ or $\| \bullet \|_q$. In particular, $\| \bullet \|_\infty$ denotes the sup-norm. In addition, we use Big O and small o as the convention.

4 Identification Results

Inspired by the proximal causal inference recently proposed by Tchetgen Tchetgen et al. [2020], we develop a non-parametric identification result for estimating $\mathcal{V}(\pi)$, which is similar to those by Bennett and Kallus [2021] and Shi et al. [2021]. Assume that we can additionally observe the so-called reward-inducing proxy variables W_t that are only related to the action A_t through (S_t, U_t) and action-inducing proxy variables Z_t that are only related to the reward R_t through (S_t, U_t) at each decision time t. See Figure 1 for a directed acyclic graph (DAG) to illustrate their relationships and a time series data example in Miao et al. [2018b]. For another example, the action-inducing proxy variables Z_t can be defined as the observed history before time t, then Z_t and related arrows in Figure 1 can be removed. Detailed assumptions and discussion are given in Appendix A. Denote the spaces of $\{Z_t\}_{t=1}^T$ and $\{W_t\}_{t=1}^T$ by $\mathcal{W}$ and $\mathcal{Z}$ respectively.

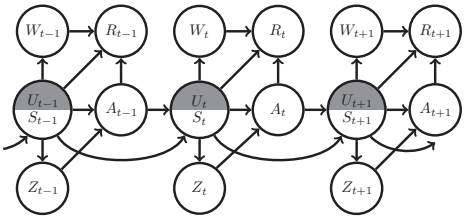


Figure 1: A representative DAG to illustrate the variables involved in the confounded POMDP.

Since the states $\{U_t\}_{t=1}^T$ are unmeasured, we cannot estimate the value function by the celebrated Bellman equation. However, with the help of confounding proxies $\{W_t, Z_t\}_{t=1}^T$, the value of a target policy π can be non-parametrically identified using observed variables under proper assumptions.

To proceed, we define a class of V-bridge functions (or V-bridges for short) $\{v_t^\pi\}_{t=1}^T$ defined over $\mathcal{W} \times \mathcal{S}$ such that for every $(s,u) \in \mathcal{S} \times \mathcal{U}$ and $t \geq 1$,

$$\mathbb{E}\left[v_t^\pi(W_t, S_t) \mid U_t = u, S_t = s\right] = \mathbb{E}^\pi\left[\textstyle\sum_{t'=t}^T R_{t'} \,\middle|\, U_t = u, S_t = s\right]. \tag{3}$$

If such V-bridges exist, then we obtain the following identification result for the policy value in (2).

Proposition 4.1 (Identification)**.** If there exist $\{v_t^\pi\}_{t=1}^T$ that satisfy (3), then the value of target policy π can be identified by $\mathcal{V}(\pi) = \mathbb{E}[v_1^\pi(W_1, S_1)]$.

Note that V-bridges $\{v_t^\pi\}_{t=1}^T$ that satisfy (3) are not necessarily unique, but we can uniquely identify $\mathcal{V}(\pi)$ based on any of them. Next, we provide a theoretical guarantee for the existence of V-bridges $\{v_t^\pi\}_{t=1}^T$ in terms of a sequence of linear integral equations.

https://doi.org/10.52202/068431-0043

Theorem 4.1. For a POMDP model of which variables satisfy the relationships illustrated in Figure 1 and some regularity conditions given in Appendix A, there always exist V-bridges $\{v_t^\pi\}_{t=1}^T$ satisfying (3). With $v_{T+1}^\pi = 0$, a particular sequence of V-bridges $\{v_t^\pi\}_{t=1}^T$ can be obtained by solving the following linear integral equations:

$$\mathbb{E}\left\{q_t^\pi(W_t, S_t, A_t) - R_t - v_{t+1}^\pi(W_{t+1}, S_{t+1}) \mid Z_t, S_t, A_t\right\} = 0, \tag{4}$$

where $\{q_t^\pi\}_{t=1}^T$ are Q-bridges defined over $\mathcal{W} \times \mathcal{S} \times \mathcal{A}$ such that

$$\mathbb{E}\left[q_t^\pi(W_t, S_t, A_t) \mid U_t = u, S_t = s, A_t = a\right] = \mathbb{E}^\pi\left[\sum_{t'=t}^T R_{t'} \,\middle|\, U_t = u, S_t = s, A_t = a\right], \tag{5}$$

for every $(s, u, a) \in \mathcal{S} \times \mathcal{U} \times \mathcal{A}$ and $t \geq 1$, and $v_t^\pi(w, s) = \sum_{a \in \mathcal{A}} \pi_t(a \mid s) q_t^\pi(w, s, a)$. Clearly Q-bridges $\{q_t^\pi\}_{t=1}^T$ also exist.

Theorem 4.1 guarantees the existence of both V-bridges and Q-bridges, and also provides a natural procedure (4) to find $\{v_t^\pi\}_{t=1}^T$ and eventually estimate the policy value $\mathcal{V}(\pi)$. Then based on Proposition 4.1 and Theorem 4.1, we can perform OPE via Algorithm 1 in the population level. Specifically at each step we will solve (4) via a non-parametric model, which is a NPIV problem.

Algorithm 1: Identification of $\mathcal{V}(\pi)$

Input: $\{(S_t, W_t, Z_t, A_t, R_t)\}_{t=1}^T$, a target policy $\pi = \{\pi_t\}_{t=1}^T$.
Let $v_{T+1}^\pi = 0$.
Repeat for $t = T, \ldots, 1$:
    Solve v_t^π and q_t^π by $\mathbb{E}\left\{q_t^\pi(W_t, S_t, A_t) - R_t - v_{t+1}^\pi(W_{t+1}, S_{t+1}) \mid Z_t, S_t, A_t\right\} = 0$ with $v_t^\pi(W_t, S_t) \triangleq \sum_{a \in \mathcal{A}} \pi_t(a \mid S_t) q_t^\pi(W_t, S_t, a)$.
Output: $\mathcal{V}(\pi) = \mathbb{E}[v_1^\pi(W_1, S_1)]$.

5 Estimation

In this section, we discuss how to estimate $\mathcal{V}(\pi)$ using batch data based on results given in Theorem 4.1 and Algorithm 1. Let a pre-collected training dataset be $\mathcal{D}_n = \{(S_{t,i}, W_{t,i}, Z_{t,i}, A_{t,i}, R_{t,i})_{t=1}^T : i = 1, \ldots, n\}$, which consists of n i.i.d. copies of the observable trajectory $(S_t, W_t, Z_t, A_t, R_t)_{t=1}^T$ of a confounded POMDP. Following Algorithm 1, we develop a FQE-type approach where we propose to solve a min-max problem for estimating v_t^π at the t-th step using the idea of Dikkala et al. [2020], and then apply Proposition 4.1 for OPE.

For convenience, we first rewrite the linear integral equations (4) for solving V-bridges in terms of operators. Define an operator $\widetilde{\mathcal{P}}_t : \mathcal{L}^2\{\mathcal{R} \times \mathcal{W} \times \mathcal{S}\} \to \mathcal{L}^2\{\mathcal{Z} \times \mathcal{S} \times \mathcal{A}\}$ such that $[\widetilde{\mathcal{P}}_t g](Z_t, S_t, A_t) = \mathbb{E}[g(R_t, W_{t+1}, S_{t+1}) \mid Z_t, S_t, A_t]$ for any $g \in \mathcal{L}^2\{\mathcal{R} \times \mathcal{W} \times \mathcal{S}\}$. Define another operator $\overline{\mathcal{P}}_t : \mathcal{L}^2\{\mathcal{W} \times \mathcal{S} \times \mathcal{A}\} \to \mathcal{L}^2\{\mathcal{Z} \times \mathcal{S} \times \mathcal{A}\}$ such that for any $h \in \mathcal{L}^2\{\mathcal{W} \times \mathcal{S} \times \mathcal{A}\}$, $[\overline{\mathcal{P}}_t h](Z_t, S_t, A_t) = \mathbb{E}[h(W_t, S_t, A_t) \mid Z_t, S_t, A_t]$. Motivated by (4), we define the *V-bridge transition operator* $\mathcal{P}_t^\pi : \mathcal{L}^2\{\mathcal{R} \times \mathcal{W} \times \mathcal{S}\} \to \mathcal{L}^2\{\mathcal{W} \times \mathcal{S}\}$ such that

$$\mathcal{P}_t^\pi g = \langle \pi_t, \mathcal{P}_t g \rangle, \quad \text{where } \mathcal{P}_t g = \overline{\mathcal{P}}_t^{-1} \widetilde{\mathcal{P}}_t g \text{ for all } g \in \mathcal{L}^2\{\mathcal{R} \times \mathcal{W} \times \mathcal{S}\}.$$

In particular, $\langle \pi_t(\cdot \mid S_t), [\mathcal{P}_t g](W_t, S_t, \cdot) \rangle \triangleq \sum_{a \in \mathcal{A}} \pi_t(a \mid S_t)[\mathcal{P}_t g](W_t, S_t, a)$, and $\widetilde{\mathcal{P}}_t g$ is invertible by $\overline{\mathcal{P}}_t$. The invertibility is ensured by Assumption 8 in Appendix A.

Then by the definition of V-bridges and (4), we can identify $\{v_t^\pi\}_{t=1}^T$ via solving

$$v_t^\pi = \mathcal{P}_t^\pi(v_{t+1}^\pi + R_t), \quad \text{for } t \geq 1. \tag{6}$$

To find the estimated V-bridges $\{\hat{v}_t^\pi\}_{t=1}^T$, it suffices to estimate $\mathcal{P}_t^\pi$. Note that one can regard (6) as a series of conditional moment model restrictions and we propose to solve them via a sequential NPIV estimation. In particular, at the t-th step, we adopt the min-max estimation method proposed by Dikkala et al. [2020] to estimate $\mathcal{P}_t^\pi$ non-parametrically as follows: $\widehat{\mathcal{P}}_t^\pi g = \left\langle \pi_t, \widehat{\mathcal{P}}_t g \right\rangle$, where

$$\widehat{\mathcal{P}}_t g/(T - t + 1) = \underset{h \in \mathcal{H}^{(t)}}{\arg\min} \left[\sup_{f \in \mathcal{F}^{(t)}} \left\{ \Psi_{t,n}(h, f, g) - \lambda(\|f\|_{\mathcal{F}^{(t)}}^2 + \frac{M}{\delta^2}\|f\|_n^2) \right\} + \lambda\mu\|h\|_{\mathcal{H}^{(t)}}^2 \right], \tag{7}$$

https://doi.org/10.52202/068431-0043

where $\|f\|_n^2 = n^{-1}\sum_{i=1}^n f^2(Z_{t,i}, S_{t,i}, A_{t,i})$ for $f \in \mathcal{F}^{(t)}$. $\mathcal{H}^{(t)}$ on $\mathcal{W}\times\mathcal{S}\times\mathcal{A}$ and $\mathcal{F}^{(t)}$ on $\mathcal{Z}\times\mathcal{S}\times\mathcal{A}$ are two user-defined function spaces endowed with norms $\|\bullet\|_{\mathcal{H}^{(t)}}$ and $\|\bullet\|_{\mathcal{F}^{(t)}}$ respectively, $\lambda, \mu, M, \delta > 0$ are tuning parameters, and

$$\Psi_{t,n}(h,f,g) = n^{-1}\sum_{i=1}^n [h(W_{t,i},S_{t,i},A_{t,i}) - (T-t+1)^{-1} g(R_{t,i},W_{t+1,i},S_{t+1,i})] f(Z_{t,i},S_{t,i},A_{t,i}),$$

where $g(R_t, W_{t+1}, S_{t+1}) = R_t + \bar{g}(W_{t+1}, S_{t+1})$ for some $\bar{g} \in \mathcal{G}^{(t+1)}$ on $\mathcal{W}\times\mathcal{S}$, endowed with norm $\|\bullet\|_{\mathcal{G}^{(t+1)}}$.

The rational behind (7) is that when $\lambda, \lambda\mu \to 0$ and $\lambda M/\delta^2 \asymp 1$, the following two population-version min-max optimization problems

$$\min_{h\in\mathcal{H}^{(t)}} \sup_{f\in\mathcal{F}^{(t)}} \mathbb{E}[h(W_t,S_t,A_t) - (T-t+1)^{-1} g(R_t,W_{t+1},S_{t+1})] f(Z_t,S_t,A_t) - \tfrac{1}{2} f^2(Z_t,S_t,A_t),$$

$$\min_{h\in\mathcal{H}^{(t)}} \mathbb{E}\{\mathbb{E}[h(W_t,S_t,A_t) - (T-t+1)^{-1} g(R_t,W_{t+1},S_{t+1}) \mid Z_t,S_t,A_t]\}^2,$$

have the same solution h when the space $\mathcal{F}^{(t)}$ of testing functions is rich enough. Note that $(T-t+1)^{-1}$ used above and in (7) are for scaling purpose.

After T steps, we output our estimator for the policy value based on the empirical counterpart of Proposition 4.1. Our FQE-type algorithm is summarized in Algorithms 2 and 3 in Appendix E.

6 Theoretical Results

In this section, we establish the finite-sample bounds for the $\mathcal{L}^2$ error of estimating V-bridge v_1^π and the error of OPE, in terms of the sample size, length of horizon and two (local) measures of ill-posedness. Our bounds also rely on the critical radii of certain spaces related to the user-defined function spaces $\mathcal{H}^{(t)}$ and $\mathcal{F}^{(t)}$ in (7), and also $\mathcal{G}^{(t)}$ of V-bridge functions.

1. Technical preliminaries. Before presenting our main results, we first introduce some concepts from the empirical process theory [Wainwright, 2019].

Definition 6.1 (Local Rademacher Complexity)**.** Given any real-valued function class $\mathcal{F}$ defined over a random vector X and any radius $\delta > 0$, the local Rademacher complexity is given by

$$\mathcal{R}_n(\mathcal{F},\delta) = \mathbb{E}_{\epsilon,X}[\sup_{f\in\mathcal{F}:\|f\|_n\le\delta} |n^{-1}\textstyle\sum_{i=1}^n \epsilon_i f(X_i)|], \tag{8}$$

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.

By bounding the local Rademacher complexity, which measures the complexity of the functional class $\mathcal{F}$ locally in a neighborhood of the ground truth, we can control the error rate of the proposed V-bridge estimator in each step. A crucial parameter for local Rademacher complexity of a function class $\mathcal{F}$ is called *critical radius*.

Definition 6.2 (Critical Radius)**.** Assume that $\mathcal{F}$ is a star-shaped function class, i.e. $\alpha f \in \mathcal{F}$ for any $f \in \mathcal{F}$ and scalar $\alpha \in [0,1]$, and also that $\mathcal{F}$ is b-uniformly bounded, i.e., $\|f\|_\infty \le b < \infty, \forall f \in \mathcal{F}$. The critical radius of $\mathcal{F}$, denoted by δ_n, is the solution to the inequality $\mathcal{R}_n(\mathcal{F},\delta) \le \delta^2/b$.

Additional Notations: We assume that the test functions f belong to a star shaped, symmetric space $\mathcal{F}^{(t)} \subseteq \mathcal{L}^2(\mathcal{Z}\times\mathcal{S}\times\mathcal{A})$ endowed with norm $\|\cdot\|_{\mathcal{F}^{(t)}}$. For brevity of notation, hereafter we suppress the time-step indicator (t) in the context unless necessary. For a function space $\mathcal{F}$, we define $\alpha\mathcal{F} = \{\alpha f : f \in \mathcal{F}\}$, for some $\alpha \in \mathbb{R}$. Define $\mathcal{F}_B = \{f \in \mathcal{F} : \|f\|_{\mathcal{F}}^2 \le B\}$, for any $B > 0$. Define the projected root mean squared error $\|\text{proj}_t f\|_2 = \sqrt{\mathbb{E}\{\mathbb{E}[f(X) \mid Z_t, S_t, A_t]\}^2}$, for any squared integrable f with respect to the conditional distribution of X given (Z_t, S_t, A_t).

Standard (Local) Measures of ill-posedness: Let $\bar{\tau}_1 = \sup_{g\in\mathcal{G}^{(1)}} \|g(W_1,S_1)\|_2/\|\mathbb{E}[g(W_1,S_1) \mid Z_1,S_1]\|_2$ be the measure of ill-posedness for $\mathcal{G}^{(1)}(\mathcal{W}_1\times\mathcal{S}_1)$ projected on $\mathcal{Z}_1\times\mathcal{S}_1$. Let $\tau_t = \sup_{h\in\mathcal{H}^{(t)}} \|h(W_t,S_t,A_t)\|_2/\|\text{proj}_t h(W_t,S_t,A_t)\|_2$ be the standard measure of ill-posedness for $\mathcal{H}^{(t)}(\mathcal{W}\times\mathcal{S}\times\mathcal{A})$ projected on $\mathcal{Z}\times\mathcal{S}\times\mathcal{A}$. It can be seen that $\bar{\tau}_1, \tau_t \ge 1$ for $t \ge 1$. Indeed we only require measuring $\bar{\tau}_1$ and τ_t *locally*. See more details in Appendix C.

2. Results. We first give Assumption 1 used to develop our theoretical results below.

Assumption 1. For each $t = 1, \ldots, T$,

(1) Closeness. For any $g \in \mathcal{G}^{(t+1)}$, $\mathcal{P}_t(g + R_t) \in \mathcal{H}^{(t)}$; For any $h \in \mathcal{H}^{(t)}$, $\langle \pi_t, h \rangle \in \mathcal{G}^{(t)}$.

(2) For any $h \in (T-t)\mathcal{H}^{(t+1)}$, we have $\|\mathcal{P}_t\left(\frac{R_t + \langle \pi_{t+1}, h\rangle}{T-t+1}\right)\|^2_{\mathcal{H}^{(t)}} \leq \|\frac{h}{T-t}\|^2_{\mathcal{H}^{(t+1)}}$.

(3) There exists a constant $C_{\mathcal{G}} > 0$ such that $\| \langle \pi_t, h \rangle \|^2_{\mathcal{G}^{(t)}} \leq C_{\mathcal{G}} \|h\|^2_{\mathcal{H}^{(t)}}$, for $h \in \mathcal{H}^{(t)}$.

(4) $q^\pi_t \in (T-t+1)\mathcal{H}^{(t)}(\mathcal{W}, \mathcal{S}, \mathcal{A})$ and $\|q^\pi_T\|^2_{\mathcal{H}^{(T)}} \leq M_{\mathcal{H}}$, where $M_{\mathcal{H}} > 0$ is a constant.

(5) Testing function class $\mathcal{F}^{(t)}$ is sufficiently rich such that there exists $L > 0$, $\|f^* - \text{proj}_t h_t\|_2 \leq \eta^{(t)}_n$, where $f^* \in \arg\min_{f \in \mathcal{F}^{(t)}_{L^2\|h_t\|^2_{\mathcal{H}}}} \|f - \text{proj}_t h_t\|_2$, for all $h_t \in \mathcal{H}^{(t)}$.

(6) Behavior policies: there exists a constant b_π such that $\pi^b_t(a \mid s) \triangleq \mathbb{E}[\tilde{\pi}^b_t(a \mid U_t, S_t) \mid S_t = s] \geq b_\pi > 0$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Assumption 1 (1) is similar to Bellman completeness, which has been widely used in RL without unobserved states [e.g., Antos et al., 2008]. Note that both $\mathcal{G}^{(t)}$ and $\mathcal{H}^{(t)}$ can be chosen as infinite-dimensional spaces, e.g., RKHSs. Hence this assumption is relatively mild. Assumption 1 (2) requires the operator $\mathcal{P}_t$ to be bounded, which can be ensured under some continuity conditions on transition kernels [Kress, 1989]. Assumption 1 (3) is a technical condition for controlling the complexity of $\mathcal{G}$ by $\mathcal{H}$. Assumption 1 (4) essentially assumes that we can model q^π (and v^π) correctly at each t-step, which is again mild as $\mathcal{H}^{(t)}$ for $t \geq 1$ can all be chosen as infinite-dimensional spaces. This assumption is also called realizability of value functions, which is commonly seen in the literature of RL [e.g., Antos et al., 2008]. Assumption 1 (5) is imposed to ensure that the space of testing functions $\mathcal{F}$ is large enough so that we are able to capture the conditional expectation operator in each min-max estimation (7). Assumption 1 (6) basically requires a full coverage of our batch data generating process induced by the behavior policy, which is widely used in OPE [Precup, 2000, Antos et al., 2008]. Next, we provide a key decomposition of the $\mathcal{L}^2$ error for V-bridge estimation.

Theorem 6.1 (Error decomposition)**.** Under Assumption 1 (1) and (6), we can decompose the $\mathcal{L}^2$ error of the estimated V-bridge by

$$\|v^\pi_1 - \hat{v}^\pi_1\|_2 \leq \bar{\tau}_1 \sum_{t=1}^T \{\Pi^t_{t'=1} C^{(t)}_{t',t'-1}\} \tau_t \|\pi_t/\pi^b_t\|_\infty \|\text{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(\hat{v}^\pi_{t+1} + R_t)\|_2,$$

where the measures of one-step transition ill-posedness $C^{(t)}_{1,0} \triangleq 1$ and $C^{(t)}_{t',t'-1}, 2 \leq t' \leq t \leq T$ are defined after Corollary 6.2.

Theorem 6.1 shows that there are four key components for upper bounding the $\mathcal{L}^2$ error of $\hat{v}^\pi_1$. The first component is the probability ratio $\|\pi_t/\pi^b_t\|_\infty$, which is used to measure the distributional mismatch between the target and behavior policies. The second component is $\|\text{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(\hat{v}^\pi_{t+1} + R_t)\|_2$, the one-step projected error of $\widehat{\mathcal{P}}_t$ to $\mathcal{P}_t$, where $\hat{v}^\pi_{t+1}$ is the estimate for v^π_{t+1} depending on the observed data after t-step. We remark that this is different from the analysis in the standard NPIV estimation with a directly measured outcome. Hence the results, e.g., from Dikkala et al. [2020], cannot be directly applied to bound this component. The last two components are related to the (local) measure of ill-posedness. The third component τ_t is the measure of ill-posedness for characterizing the difficulty of estimating q^π_t by (4) using $\mathcal{H}^{(t)}$ at the t-th step. $\{\tau_t\}^T_{t=1}$ are similar to those used in the standard NPIV estimation such as Chen and Reiss [2011], and the effect of each τ_t on the upper bound is cumulative. The last component $\{\Pi^t_{t'=1} C^{(t)}_{t',t'-1}\}^T_{t=1}$ quantify the propagation effect of estimation errors in previous steps on the last step of estimating v^π_1, which is multiplicative in terms of $C^{(t)}_{t',t'-1}$. We call $C^{(t)}_{t',t'-1}$ the measure of one-step transition ill-posedness from t' to $t'-1$ related to t-step NPIV estimations. Next we provide detailed bounds for the second and last components. The discussion of the third component can be found in Appendix C.4.

Component 2: one-step projected error. In the following, we show that $\|\text{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(\hat{v}^\pi_{t+1} + R_t)\|_2$ is bounded by the critical radii of some spaces defined as balls $\mathcal{H}^{(t)}_B$, $\mathcal{G}^{(t+1)}_{C_G(T-t+1)M_{\mathcal{H}}}$ in hypothesis spaces $\mathcal{H}^{(t)}$, $\mathcal{G}^{(t+1)}$ respectively and a ball $\mathcal{F}^{(t)}_{3M}$ in testing space $\mathcal{F}^{(t)}$, for some fixed constants $M, B > 0$ such that functions in $\mathcal{H}^{(t)}_B$ and $\mathcal{F}^{(t)}_{3M}$ have uniformly bounded ranges in $[-1, 1]$ for all $1 \leq t \leq T$. Let

$$\mathbf{\Omega}^{(t)} = \{(s_t, w_t, z_t, a_t, s_{t+1}, w_{t+1}) \mapsto r(h^*_g(w_t, s_t, a_t) - g(w_{t+1}, s_{t+1}))f(z_t, s_t, a_t) : g \in \mathcal{G}^{(t+1)}_{C_G(T-t+1)M_{\mathcal{H}}}, f \in \mathcal{F}^{(t)}_{3M}, r \in [0, 1]\}, \text{ and}$$

https://doi.org/10.52202/068431-0043

$$\Xi^{(t)} = \{(s_t, w_t, z_t, a_t) \mapsto r[h - h_g^*](w_t, s_t, a_t) f^{L^2 B}(z_t, s_t, a_t) : \\ h \in \mathcal{H}^{(t)}, h - h_g^* \in \mathcal{H}_B^{(t)}, g \in \mathcal{G}^{(t+1)}_{C_G(T-t+1)M_{\mathcal{H}}}, r \in [0,1]\},$$

where $h_g^* \in \mathcal{H}^{(t)}$ is the solution to $\mathbb{E}\left[h(W_t, S_t, A_t) - g(W_{t+1}, S_{t+1}) \mid Z_t, S_t, A_t\right] = 0$, and $f^{L^2 B} = \arg\min_{f \in \mathcal{F}^{(t)}_{L^2 B}} \|f - \text{proj}_t(h - h_g^*)\|_2$ for a given $L > 0$. An upper bound for $\|\text{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(\hat{v}^{\pi}_{t+1} + R_t)\|_2$ is given in Theorem 6.2.

Theorem 6.2. Suppose that Assumption 1 holds. Let $\delta_n^{(t)} = \bar{\delta}_n^{(t)} + c_0\sqrt{\frac{\log(c_1 T/\zeta)}{n}}$ for some universal constants $c_0, c_1 > 0$ where $\bar{\delta}_n^{(t)}$ is the upper bound of the critical radii of $\mathcal{F}_{3M}^{(t)}$, $\mathbf{\Omega}^{(t)}$ and $\mathbf{\Xi}^{(t)}$. Assume that the approximation error in Assumption 1 (5) can be bounded by $\eta_n^{(t)} \le \delta_n^{(t)}$. Furthermore, letting tuning parameters satisfy $M\lambda \asymp (\delta_n^{(t)})^2$ and $\mu \ge \mathcal{O}(L^2 + M/B)$, with probability at least $1 - \zeta$, we have

$$\|\text{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(\hat{v}^{\pi}_{t+1} + R_t)\|_2 \lesssim M_{\mathcal{H}}(T - t + 1)^2 \delta_n^{(t)} \quad \text{for all } 1 \le t \le T.$$

Depending on the choices of $\mathcal{H}^{(t)}$, $\mathcal{G}^{(t+1)}$, and $\mathcal{F}^{(t)}$, we can obtain different finite-sample error bounds of the one-step projected error for each t. Below we provide two examples.

Corollary 6.1. Let $\mathcal{F}^{(t)}$, $\mathcal{H}^{(t)}$ and $\mathcal{G}^{(t+1)}$ be VC-subgraph classes with VC dimensions $\mathbb{V}(\mathcal{F}^{(t)})$, $\mathbb{V}(\mathcal{H}^{(t)})$ and $\mathbb{V}(\mathcal{G}^{(t+1)})$ respectively. Then with probability at least $1 - \zeta$, for all $1 \le t \le T$,

$$\|\text{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(\hat{v}^{\pi}_{t+1} + R_t)\|_2 \lesssim (T - t + 1)^{2.5}\sqrt{\frac{\log(c_1 T/\zeta)\max\{\mathbb{V}(\mathcal{F}^{(t)}), \mathbb{V}(\mathcal{H}^{(t)}), \mathbb{V}(\mathcal{G}^{(t+1)})\}}{n}}.$$

The definition of the VC-subgraph class can be found in, e.g., Wainwright [2019]. This is a broad class. For example, if one lets each of $\mathcal{F}^{(t)}$, $\mathcal{H}^{(t)}$ and $\mathcal{G}^{(t+1)}$ be a linear space $\mathcal{F} = \{\theta^{\top}\phi(\cdot) : \theta \in \mathbb{R}^d\}$ with basis functions $\phi(\cdot)$, then $\mathbb{V}(\mathcal{F}) = d + 1$. Then the upper bound for the one-step projected error becomes $\mathcal{O}((T - t + 1)^{2.5} d/\sqrt{n})$.

Corollary 6.2. Let $\mathcal{H}^{(t)}$, $\mathcal{G}^{(t+1)}$ and $\mathcal{F}^{(t)}$ be reproducing kernel Hilbert spaces (RKHSs) equipped with kernels $K_{\mathcal{H}^{(t)}}$, $K_{\mathcal{G}^{(t+1)}}$ and $K_{\mathcal{F}^{(t)}}$ respectively. For a given positive definite kernel K, we denote its nonincreasing eigenvalue sequence by $\{\lambda_j^{\downarrow}(K)\}_{j=1}^{\infty}$. We consider two scenarios for $\{\lambda_j^{\downarrow}(K)\}_{j=1}^{\infty}$.

(1) **Polynomial eigen-decay**: If $\lambda_j^{\downarrow}(K_{\mathcal{H}^{(t)}}) \le a j^{-2\alpha_{\mathcal{H}}}$, $\lambda_j^{\downarrow}(K_{\mathcal{G}^{(t+1)}}) \le a j^{-2\alpha_{\mathcal{G}}}$ and $\lambda_j^{\downarrow}(K_{\mathcal{F}^{(t)}}) \le a j^{-2\alpha_{\mathcal{F}}}$ for constants $\alpha_{\mathcal{H}}, \alpha_{\mathcal{G}}, \alpha_{\mathcal{F}} > 1/2$ and $a > 0$, then under the assumptions in Theorem 6.2, with probability at least $1 - \zeta$, for all $1 \le t \le T$, we have

$$\|\text{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(\hat{v}^{\pi}_{t+1} + R_t)\|_2 \lesssim (T - t + 1)^{2.5}\sqrt{\log(c_1 T/\zeta)} n^{-\frac{1}{2 + \max\{1/\alpha_{\mathcal{H}}, 1/\alpha_{\mathcal{G}}, 1/\alpha_{\mathcal{F}}\}}} \log(n).$$

(2) **Exponential eigen-decay**: If $\lambda_j^{\downarrow}(K_{\mathcal{H}^{(t)}}) \le a_1 e^{-a_2 j^{\beta_{\mathcal{H}}}}$, $\lambda_j^{\downarrow}(K_{\mathcal{G}^{(t+1)}}) \le a_1 e^{-a_2 j^{\beta_{\mathcal{G}}}}$ and $\lambda_j^{\downarrow}(K_{\mathcal{F}^{(t)}}) \le a_1 e^{-a_2 j^{\beta_{\mathcal{F}}}}$, for constants $a_1, a_2, \beta_{\mathcal{H}}, \beta_{\mathcal{G}}, \beta_{\mathcal{F}} > 0$, then under the assumptions in Theorem 6.2, with probability at least $1 - \zeta$, for all $1 \le t \le T$, we have

$$\|\text{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(\hat{v}^{\pi}_{t+1} + R_t)\|_2 \lesssim (T - t + 1)^{2.5}\left\{\sqrt{\frac{(\log n)^{1/\min\{\beta_{\mathcal{H}}, \beta_{\mathcal{G}}, \beta_{\mathcal{F}}\}}}{n}} + \sqrt{\frac{\log(c_1 T/\zeta)}{n}}\right\}.$$

Kernels of the two types of eigen-decay considered above are very common. For example, the kernel of the α-order Soblev space with $\alpha > 1/2$, has a polynomial eigen-decay while the Gaussian kernel has an exponential eigen-decay, with $\beta = 2$ for Lebesgue measure on real line and $\beta = 1$ on a compact domain [Wei et al., 2017].

Components 4: measure of one-step transition ill-posedness. We first provide more insights on $\{\Pi_{t'=1}^{t} C_{t',t'-1}^{(t)}\}_{t=1}^{T}$ before providing an upper bound. We formally define the local measure of one-step transition ill-posedness $C_{t'+1,t'}^{(t)}$ recursively based on $C_{t,t-1}^{(t)}$ to $C_{2,1}^{(t)}$ as

$$C_{t'+1,t'}^{(t)} \triangleq \sup_{g \in \mathcal{G}(W_{t'+1} \times S_{t'+1})} \frac{\|\mathbb{E}^{\pi_{t'}}[g(W_{t'+1}, S_{t'+1}) \mid Z_{t'}, S_{t'}]\|_2}{\|\mathbb{E}[g(W_{t'+1}, S_{t'+1}) \mid Z_{t'+1}, S_{t'+1}]\|_2}, \quad \text{subject to}$$

$$\|\mathbb{E}[g(W_{t'+1}, S_{t'+1}) \mid Z_{t'+1}, S_{t'+1}]\|_2 \lesssim \tau_t (T - t + 1)^2 \delta_n^{(t)} \|\pi_{t'}/\pi_{t'}^b\|_{\infty} \textstyle\prod_{s=t'+1}^{t-1} C_{s+1,s}^{(t)},$$

with $C^{(t)}_{t+1,t} \triangleq 1$ for each $t = 1, \ldots, T$. For $C^{(t)}_{t,t-1}$, we can upper bound $\|\mathbb{E}[\hat{v}^{\pi}(W_t, S_t) \mid Z_t, S_t]\|_2$ by the projected error $\|\text{proj}_t(\widehat{\mathcal{P}}_t - \mathcal{P}_t)(\hat{v}^{\pi}_{t+1} + R_t)\|_2$ multiplied by the ill-posedness τ_t. By Theorem 6.2, the projected error can be well controlled by $\delta^{(t)}_n$ with high probability, so $C^{(t)}_{t,t-1}$ can be defined locally. Therefore, we can provide an upper bound for the denominator sequentially and define all $C^{(t)}_{t'+1,t'}$ locally, which indicates that all $C^{(t)}_{t'+1,t'}$ could be small.

For example, if we use the observed history as the action-inducing proxy, then $\sigma(\mathcal{Z}_1 \times \mathcal{S}_1) \subset \sigma(\mathcal{Z}_2 \times \mathcal{S}_2) \subset \cdots \subset \sigma(\mathcal{Z}_T \times \mathcal{S}_T)$ is a filtration. In this case, $\prod_{t'=2}^{t} C^{(t)}_{t',t'-1}$ are expected to be small for $t \geq 2$ if the target policy is stationary. While this can enlarge the critical radii $\delta^{(t)}_n$ due to the dimension of the action-inducing proxy, this only affects one-step errors. See detailed discussion in Appendix B. Motivated by this, it is reasonable to impose Assumption 2 below on $C^{(t)}_{t'+1,t'}$.

Assumption 2. For every $t \geq 2$ and $2 \leq m \leq t$, $C^{(t)}_{m,m-1} \leq 1 + \frac{a_t}{m^{\alpha_t}}$ with time-dependent constants $a_t > 0, \alpha_t \geq \alpha > 1$.

Corollary 6.3. If Assumption 2 holds, then $\prod_{t'=1}^{t} C^{(t)}_{t',t'-1} \leq \exp\{a_t \zeta(\alpha_t)\}$, where $\zeta(\alpha_t) = \sum_{n=1}^{\infty}(1/n)^{\alpha_t}$ is uniformly bounded for $t \geq 1$.

Main result: error bounds for V-bridge estimation and OPE. Define $\texttt{trans-ill} = \max_{1\leq t\leq T} \exp\{a_t\zeta(\alpha_t)\}$ and let $\texttt{ill}_{\max} = \bar{\tau}_1 \max_{1\leq t\leq T} \tau_t \|\pi_t/\pi^b_t\|_{\infty}$. Summarizing all aforementioned results, we have the following main theorem based on the polynomial eigen-decay case in Corollary 6.2. Other cases can be found in Appendix B.

Theorem 6.3 (Finite-sample error bounds for V-bridges and policy value)**.** Under Assumptions 1 and 2, and assumptions in Theorem 6.2 and Corollary 6.2 (1), with probability at least $1-\zeta$, we have

$$\|v^{\pi}_1 - \hat{v}^{\pi}_1\|_2 \lesssim \texttt{ill}_{\max} \times \texttt{trans-ill} \times T^{7/2}\sqrt{\log(c_1 T/\zeta)} n^{-\frac{1}{2+\max\{1/\alpha_{\mathcal{H}}, 1/\alpha_{\mathcal{G}}, 1/\alpha_{\mathcal{F}}\}}} \log(n), \text{ and}$$

$$|\mathcal{V}(\pi) - \hat{\mathcal{V}}(\pi)| \lesssim \texttt{ill}_{\max} \times \texttt{trans-ill} \times T^{7/2}\sqrt{\log(c_1 T/\zeta)} n^{-\frac{1}{2+\max\{1/\alpha_{\mathcal{H}}, 1/\alpha_{\mathcal{G}}, 1/\alpha_{\mathcal{F}}\}}} \log(n).$$

Theorem 6.3 provides the first finite-sample error bound for OPE under confounded and episodic POMDPs in terms of the sample size, length of horizon and two (*local*) measures of ill-posedness. Without considering the measures of ill-posedness, the derived error bound for V-bridge function nearly achieves the optimal $\mathcal{L}^2$-convergence rate in the classical non-parametric regression [Stone, 1982]. Moreover, our OPE error bound depends on a polynomial order of T, i.e., $T^{7/2}$, which is larger than the standard $\mathcal{O}(T^3)$ in the OPE without unobserved variables. However, when the function class consider in (7) grows with the sample size n, $\texttt{ill}_{\max}$ will also increase and therefore the convergence rates in Theorem 6.3 could be much slower. Next we study a case when we can control the local measures of ill-posedness $\{\tau_t\}_{t=1}^{T}$, by assuming that $\lambda_{\min}(\Gamma^{(t)}_m) \geq \nu_m$ for all $1 \leq t \leq T$ almost surely and other regularity conditions in Lemma C.1, where $\Gamma^{(t)}_m \triangleq \mathbb{E}\left\{\mathbb{E}[e^{(t)}_I(W_t, S_t, A_t) \mid Z_t, S_t, A_t]\mathbb{E}[e^{(t)}_I(W_t, S_t, A_t) \mid Z_t, S_t, A_t]^{\top}\right\}$ with $e^{(t)}_I = (e^{(t)}_1, \ldots, e^{(t)}_m)$ as the first m eigenfunctions of kernel $K_{\mathcal{H}^{(t)}}$. Similar conditions can be imposed to control $\bar{\tau}_1$, which is omitted here for simplicity. Let $\eta(n, T, \zeta, \alpha_{\mathcal{H}}, \alpha_{\mathcal{F}}, \alpha_{\mathcal{G}}, b) \triangleq T^{\frac{7(\alpha_{\mathcal{H}}-1/2)+10b}{2(\alpha_{\mathcal{H}}-1/2)+4b}}\left(\sqrt{\log(c_1 T/\zeta)} n^{-\frac{1}{2+\max\{1/\alpha_{\mathcal{H}}, 1/\alpha_{\mathcal{G}}, 1/\alpha_{\mathcal{F}}\}}} \log(n)\right)^{\frac{\alpha_{\mathcal{H}}-1/2}{\alpha_{\mathcal{H}}-1/2+2b}}$, with b defined below.

Corollary 6.4. If assumptions in Theorem 6.3 holds and $\nu_m \geq m^{-2b}$ for some $b \geq 0$, then

$$\|v^{\pi}_1 - \hat{v}^{\pi}_1\|_2 \lesssim \bar{\tau}_1 \max_{1\leq t\leq T} \|\pi_t/\pi^b_t\|_{\infty} \times \texttt{trans-ill} \times \eta(n, T, \zeta, \alpha_{\mathcal{H}}, \alpha_{\mathcal{F}}, \alpha_{\mathcal{G}}, b),$$

$$|\mathcal{V}(\pi) - \hat{\mathcal{V}}(\pi)| \lesssim \bar{\tau}_1 \max_{1\leq t\leq T} \|\pi_t/\pi^b_t\|_{\infty} \times \texttt{trans-ill} \times \eta(n, T, \zeta, \alpha_{\mathcal{H}}, \alpha_{\mathcal{F}}, \alpha_{\mathcal{G}}, b).$$

Corollary C.5.2 considers the mildly ill-posed case, i.e., $\nu_m \geq m^{-2b}$, and shows that the local measure of ill-posedness can deteriorate convergence rate of $\hat{v}^{\pi}_1$ significantly. If b is large relative to $\alpha_{\mathcal{H}}$ or further severely ill-posed case is considered (i.e., ν_m decays exponentially fast, see Appendix C.5.2), then the convergence rate of V-bridge estimation could be much slower and the typical requirement on the nuisance parameter for achieving asymptotic normality for the policy value will fail. On the other hand, it can be seen that when $b = 0$, the finite sample error bounds match the results in Theorem 6.3.

7 Simulation

In this section, we perform a simulation study to evaluate the performance of our proposed OPE estimation and to verify the finite-sample error bound of our OPE estimator in Theorem 6.3.

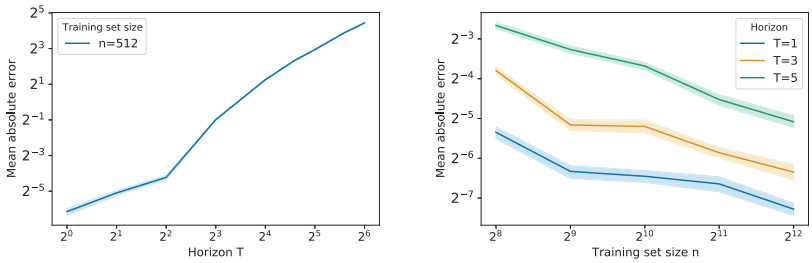
https://doi.org/10.52202/068431-0043

Let $\mathcal{S} = \mathbb{R}^2, \mathcal{U} = \mathbb{R}, \mathcal{W} = \mathbb{R}, \mathcal{Z} = \mathbb{R}$, and $\mathcal{A} = \{1, -1\}$. At time t, the hidden state U_t, two proximal variables Z_t, W_t satisfy the following multivariate normal distribution given (S_t, A_t):

$$(Z_t, W_t, U_t) \mid (S_t, A_t) \sim \mathcal{N}\left(\begin{bmatrix} \alpha_0 + \alpha_a A_t + \alpha_s S_t \\ \mu_0 + \mu_a A_t + \mu_s S_t \\ \kappa_0 + \kappa_a A_t + \kappa_s S_t \end{bmatrix}, \Sigma = \begin{bmatrix} \sigma_z^2 & \sigma_{zw} & \sigma_{zu} \\ \sigma_{zw} & \sigma_w^2 & \sigma_{wu} \\ \sigma_{zu} & \sigma_{wu} & \sigma_u^2 \end{bmatrix} \right), \tag{9}$$

where parameters are given in the Appendix.

The behavior policy is given by $\tilde{\pi}_t^b(A_t \mid U_t, S_t) = \text{expit}\left\{-A_t\left(t_0 + t_u U_t + t_s^\top S_t\right)\right\}$, where $t_0 = 0$, $t_u = 1$, and $t_s^\top = [-0.5, -0.5]$. Then by Assumption 1 (6), $\pi_t^b(A_t \mid S_t) = \text{expit}\{-A_t\left(t_0 + t_u\kappa_0 + (t_s + t_u\kappa_s)^\top S_t\right)\}$. The initial S_1 is uniformly sampled from $\mathbb{R}^2$. At time t, given (S_t, U_t, A_t), we generate $S_{t+1} = S_t + A_t U_t \mathbf{1}_2 + e_{S_{t+1}}$, where $\mathbf{1}_2 = [1, 1]^\top$ and the random error $e_{S_{t+1}} \sim \mathcal{N}([0, 0]^\top, \mathbf{I}_2)$ with $\mathbf{I}_2$ denoting the 2-by-2 identity matrix. The reward is given by $R_t = \text{expit}\left\{\frac{1}{2} A_t(U_t + [1, -2]S_t)\right\} + e_t$, where $e_t \sim \text{Uniform}[-0.1, 0.1]$. One can verify that our simulation setting satisfies the conditions in Section A.1 so that our method can be applied.



(a) (b)

Figure 2: Simulation results for OPE errors $|\widehat{\mathcal{V}}(\pi) - \mathcal{V}(\pi)|$. Mean absolute errors (solid lines) and their standard error bands (shaded regions) are displayed for different combinations of (n, T).

We choose $\mathcal{F}^{(t)}$ and $\mathcal{H}^{(t)}$ as RKHSs endowed with Gaussian kernels, with bandwidths selected according to the median heuristic trick by Fukumizu et al. [2009] for each $1 \leq t \leq T$. The pool of scaling factors SCALE contains 30 positive numbers spaced evenly on a log scale between 0.001 to 0.05. The number of cross-validation partition $K = 5$. The true target policy value of π is estimated by the mean cumulative rewards of $50,000$ Monte Carlo trajectories with policy π. We compare our OPE estimator $\widehat{\mathcal{V}}(\pi)$ with the target policy value by computing mean absolute error (MAE) for each setting of (n, T), as reported in Figure 2. Figure 2 validate the derived finite-sample error bound of our OPE estimator in Theorem 6.3. Specifically, Figure 2 (a) shows that the OPE estimation error is polynomial in T, but with an order slightly smaller than $\mathcal{O}(T^{7/2})$ as stated in Theorem 6.3. Figure 2 (b) shows that the convergence rate in terms of the sample size n for our OPE estimator is slower than $\mathcal{O}(n^{-1/2})$, which also justifies our theoretical results.

8 Discussion

In this paper, we propose a non-parametric identification and estimation method for OPE in episodic confounded POMDPs with continuous states, relying on time-dependent proxy variables. We develop a fitted-Q-evaluation-type algorithm for estimating the V-bridge functions sequentially and for OPE based on the estimated V-bridges. The first finite-sample error bound for estimating the policy value under confounded POMDPs is established, which achieves a polynomial order with respect to the sample size and the length of horizon. Our OPE results can serve as a foundation for developing new policy optimization algorithms in the confounded POMDP, which We will leave for future work.

Acknowledgement

Zhang's research is partially supported by GW University Facilitating Fund.

References

C. Ai and X. Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

C. Ai and X. Chen. The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions. *Journal of Econometrics*, 170(2):442–457, 2012.

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

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

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

R. Blundell, X. Chen, and D. Kristensen. Semi-nonparametric IV estimation of shape-invariant Engel curves. *Econometrica*, 75(6):1613–1669, 2007.

D. A. Bruns-Smith. Model-free and model-based policy evaluation when causality is uncertain. In *International Conference on Machine Learning*, pages 1116–1126. PMLR, 2021.

Q. Cai, Z. Yang, and Z. Wang. Sample-efficient reinforcement learning for pomdps with linear function approximations. *arXiv preprint arXiv:2204.09787*, 2022.

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

X. Chen and T. M. Christensen. Optimal sup-norm rates and uniform inference on nonlinear functionals of nonparametric iv regression. *Quantitative Economics*, 9(1):39–84, 2018.

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

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

X. Chen, V. Chernozhukov, S. Lee, and W. K. Newey. Local identification of nonparametric and semiparametric models. *Econometrica*, 82(2):785–809, 2014.

S. Darolles, Y. Fan, J.-P. Florens, and E. Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011.

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

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

X. D'Haultfoeuille. On the completeness condition in nonparametric instrumental problems. *Econometric Theory*, 27(3):460–471, 2011.

D. J. Foster and V. Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

K. Fukumizu, A. Gretton, G. R. Lanckriet, B. Schölkopf, and B. K. Sriperumbudur. Kernel choice and classifiability for RKHS embeddings of probability distributions. In *Advances in Neural Information Processing Systems*, pages 1750–1758, 2009.

P. Hall and J. L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics*, 33(6):2904–2929, 2005.

E. A. Hansen, D. S. Bernstein, and S. Zilberstein. Dynamic programming for partially observable stochastic games. In *AAAI*, volume 4, pages 709–715, 2004.

https://doi.org/10.52202/068431-0043

J. Hartford, G. Lewis, K. Leyton-Brown, and M. Taddy. Deep IV: A flexible approach for counterfactual prediction. In *International Conference on Machine Learning*, pages 1414–1423. PMLR, 2017.

A. Hinrichs. Optimal Weyl inequality in Banach spaces. *Proceedings of the American Mathematical Society*, 134(3):731–735, 2006.

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

C. Jin, S. Kakade, A. Krishnamurthy, and Q. Liu. Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems*, 33:18530–18539, 2020.

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

R. Kress. *Linear Integral Equations*, volume 82. Springer, 1989.

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

J. Li, Y. Luo, and X. Zhang. Causal reinforcement learning: An instrumental variable approach. *arXiv preprint arXiv:2103.04021*, 2021.

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

M. Littman and R. S. Sutton. Predictive representations of state. *Advances in Neural Information Processing Systems*, 14, 2001.

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

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

K. Muandet, A. Mehrjou, S. K. Lee, and A. Raj. Dual instrumental variable regression. *Advances in Neural Information Processing Systems*, 33:2710–2721, 2020.

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

H. Namkoong, R. Keramati, S. Yadlowsky, and E. Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33:18819–18831, 2020.

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

C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of markov decision processes. *Mathematics of Operations Research*, 12(3):441–450, 1987.

A. Pietsch. Eigenvalues and s-numbers. *Cambridge Studies in Advanced Mathematics*, 13, 1987.

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

A. N. Rafferty, E. Brunskill, T. L. Griffiths, and P. Shafto. Faster teaching by POMDP planning. In *International Conference on Artificial Intelligence in Education*, pages 280–287. Springer, 2011.

C. Shi, M. Uehara, and N. Jiang. A minimax learning approach to off-policy evaluation in partially observable Markov decision processes. *arXiv preprint arXiv:2111.06784*, 2021.

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

X. Shi, W. Miao, J. C. Nelson, and E. J. Tchetgen Tchetgen. Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(2):521–540, 2020.

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

S. Singh, M. James, and M. Rudary. Predictive state representations: A new theory for modeling dynamical systems. *arXiv preprint arXiv:1207.4167*, 2012.

C. J. Stone. Optimal global rates of convergence for nonparametric regression. *Annals of Statistics*, pages 1040–1053, 1982.

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

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

A. Tsoukalas, T. Albertson, and I. Tagkopoulos. From data to optimal decision making: a data-driven, probabilistic machine learning approach to decision support for patients with sepsis. *JMIR Medical Informatics*, 3(1):e3445, 2015.

M. J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Y. Wei, F. Yang, and M. J. Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. *Advances in Neural Information Processing Systems*, 30, 2017.

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

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

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] The paper's contributions and scope are summarized in the abstract and then emphasized in Sections 1 and 2.
   (b) Did you describe the limitations of your work? [Yes] We did not provide results of minimax lower bound for our estimation problem.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] Main assumptions are stated in Sections 4 and 6. Some regularity conditions are provided in Appendix A.
   (b) Did you include complete proofs of all theoretical results? [Yes] All proofs for theoretical results are provided in Appendices C and D.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We attached the code and instructions in the supplemental materials.

https://doi.org/10.52202/068431-0043

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Provided in Appendix F

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [Yes]

(b) Did you mention the license of the assets? [Yes]

(c) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(d) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

https://doi.org/10.52202/068431-0043