PESSIMISM IN THE FACE OF CONFOUNDERS: PROVABLY EFFICIENT OFFLINE REINFORCEMENT LEARNING IN PARTIALLY OBSERVABLE MARKOV DECISION PROCESSES

Anonymous authors
Paper under double-blind review

ABSTRACT

We study offline reinforcement learning (RL) in partially observable Markov decision processes. In particular, we aim to learn an optimal policy from a dataset collected by a behavior policy which possibly depends on the latent state. Such a dataset is confounded in the sense that the latent state simultaneously affects the action and the observation, which is prohibitive for existing offline RL algorithms. To this end, we propose the Proxy variable Pessimistic Policy Optimization (`P3O`) algorithm, which addresses the confounding bias and the distributional shift between the optimal and behavior policies in the context of general function approximation. At the core of `P3O` is a coupled sequence of pessimistic confidence regions constructed via proximal causal inference, which is formulated as minimax estimation. Under a partial coverage assumption on the confounded dataset, we prove that `P3O` achieves a $n^{-1/2}$-suboptimality, where n is the number of trajectories in the dataset. To our best knowledge, `P3O` is the first provably efficient offline RL algorithm for POMDPs with a confounded dataset.

1 INTRODUCTION

Offline reinforcement learning (RL) (Sutton and Barto, 2018) aims to learn an optimal policy of a sequential decision making problem purely from an offline dataset collected a priori, without any further interactions with the environment. Offline RL is particularly pertinent to applications in critical domains such as precision medicine (Gottesman et al., 2019) and autonomous driving (Shalev-Shwartz et al., 2016). In particular, in these scenarios, interacting with the environment via online experiments might be risky, slow, or even possibly unethical. But oftentimes offline datasets consisting of past interactions, e.g., treatment records for precision medicine (Chakraborty and Moodie, 2013; Chakraborty and Murphy, 2014) and human driving data for autonomous driving (Sun et al., 2020), are adequately available. As a result, offline RL has attracted substantial research interest recently (Levine et al., 2020).

Most of the existing works on offline RL develop algorithms and theory on the model of Markov decision processes (MDPs). However, in many real-world applications, due to certain privacy concerns or limitations of the sensor apparatus, the states of the environment cannot be directly stored in the offline datasets. Instead, only partial observations generated from the states of the environments are stored (Dulac-Arnold et al., 2021). For example, in precision medicine, a physician's treatment might consciously or subconsciously depend on the patient's mood and socioeconomic status (Zhang and Bareinboim, 2016), which are not recorded in the data due to privacy concerns. As another example, in autonomous driving, a human driver makes decisions based on multimodal information of the environment that is not limited to visual and auditory inputs, but only observations captured by LIDARs and cameras are stored in the datasets (Sun et al., 2020). In light of the partial observations in the datasets, these situations are better modeled as partially observable Markov decision processes (POMDPs) (Lovejoy, 1991). Existing offline RL methods for MDPs, which fail to handle partial observations, are thus not applicable.

In this work, we make the initial step towards studying offline RL in POMDPs where the datasets only contain partial observations of the states. In particular, motivated from the aforementioned

real-world applications, we consider the case where the behavior policy takes actions based on the states of the environment, which are not part of the dataset and thus are latent variables. Instead, the trajectories in datasets consist of partial observations emitted from the latent states, as well as the actions and rewards. For such a dataset, our goal is to learn an optimal policy in the context of general function approximation.

Furthermore, offline RL in POMDP suffers from several challenges. First of all, it is known that both planning and estimation in POMDPs are intractable in the worst case (Papadimitriou and Tsitsiklis, 1987; Burago et al., 1996; Goldsmith and Mundhenk, 1998; Mundhenk et al., 2000; Vlassis et al., 2012). Thus, we have to identify a set of sufficient conditions that warrants efficient offline RL. More importantly, our problem faces the unique challenge of the confounding issue caused by the latent states, which does not appear in either online and offline MDPs or online POMDPs. In particular, both the actions and observations in the offline dataset depend on the unobserved latent states, and thus are confounded (Pearl, 2009).

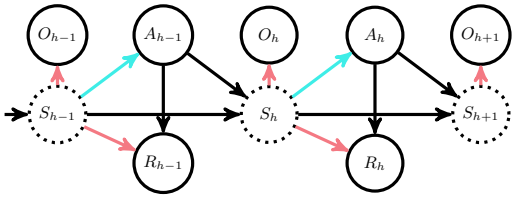

Figure 1: Causal graph of the data generating process for offline RL in POMDP. Here S_h is the state at step h. Besides, A_h, R_h, and O_h are the action, immediate reward, and observation, respectively. The dotted nodes indicate that the variables are not stored in the offline dataset. Solid arrows indicate the dependency among the variables. In specific, the action A_h is specified by the behavior policy which is a function of S_h (**blue** arrows). Moreover, both the observation O_h and reward R_h depend on the state S_h (**red** arrows). We remark that S_h affects both A_h and (O_h, R_h) and thus serves as an **unobserved confounder**.

Such a confounding issue is illustrated by a causal graph in Figure 1. As a result, directly applying offline RL methods for MDPs will nevertheless incur a considerable confounding bias. Besides, since the latent states evolve according to the Markov transition kernel, the causal structure is thus dynamic, which makes the confounding issue more challenging to handle than that in static causal problems. Furthermore, apart from the confounding issue, since we aim to learn the optimal policy, our algorithm also needs to handle the distributional shift between the trajectories induced by the behavior policy and the family of target policies. Finally, to handle large observation spaces, we need to employ powerful function approximators. As a result, the coupled challenges due to (i) the confounding bias, (ii) distributional shift, and (iii) large observation spaces that are distinctive in our problem necessitates new algorithm design and theory.

To this end, by leveraging tools from proximal causal inference (Lipsitch et al., 2010; Tchetgen et al., 2020; Miao et al., 2018b;a), we propose the Proxy variable Pessimistic Policy Optimization (`P3O`) algorithm, which provably addresses the challenge of the confounding bias and the distributional shift in the context of general function approximation. Specifically, we focus on a benign class of POMDPs where the causal structure involving latent states can be captured by the past and current observations, which serves as the negative control action and outcome respectively (Miao et al., 2018a;b; Cui et al., 2020; Singh, 2020; Kallus et al., 2021; Bennett and Kallus, 2021; Shi et al., 2021). Then the value of each policy can be identified by a set of confounding bridge functions corresponding to that policy, which satisfy a sequence of backward moment equations that are similar to the celebrated Bellman equations in classical RL (Bellman and Kalaba, 1965). Thus, by estimating these confounding bridge functions from offline data, we can estimate the value of each policy without incurring the confounding bias.

More concretely, `P3O` involves two components — policy evaluation via minimax estimation and policy optimization via pessimism. Specifically, to tackle the distributional shift, `P3O` returns the policy that maximizes pessimistic estimates of the values obtained by policy evaluation. Meanwhile, in policy evaluation, to ensure pessimism, we construct a coupled sequence of confidence regions for the confounding bridge functions via minimax estimation, using function approximators. Furthermore, under a partial coverage assumption on the confounded dataset, we prove that `P3O` achieves a $\widetilde{\mathcal{O}}(H\sqrt{\log(\mathcal{N}_{\text{fun}})/n})$ suboptimality, where n is the number of trajectories, H is the length of each trajectory, $\mathcal{N}_{\text{fun}}$ stands for the complexity of the employed function classes (e.g., the covering number), and $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors. When specified to linear function classes, the suboptimality of `P3O` becomes $\widetilde{\mathcal{O}}(\sqrt{H^3 d/n})$, where d is the dimension of the feature mapping. To our best knowledge, we establish the first provably efficient offline RL algorithm for POMDP with a confounded dataset.

	Offline	Partial Obs.	Confounded Data	Policy Opt.
Xie et al. (2021)	✓	✘	✘	✓
Uehara and Sun (2021)	✓	✘	✘	✓
Jin et al. (2020)	✘	✓	✘	✓
Efroni et al. (2022)	✘	✓	✘	✓
Liu et al. (2022)	✘	✓	✘	✓
Bennett and Kallus (2021)	✓	✓	✓	✘
Shi et al. (2021)	✓	✓	✓	✘
P3O (ours)	✓	✓	✓	✓

Table 1: We compare with most related representative works in closely related lines of research. The first line of research studies offline RL in standard MDPs without any partial observability. The second line of research studies online RL in POMDPs where the actions are specified by history-dependent policies. Thus, the actions does not directly depends on the latent states and thus these works do not involve the challenge due to confounded data. The third line of research studies OPE in POMDPs where the goal is to learn the value of the target policy as opposed to learning the optimal policy. As a result, these works do not to need to handle the challenge of distributional shift via pessimism.

1.1 Overview of Techniques

To deal with the coupled challenges of confounding bias, distributional shift, and large observational spaces, our algorithm and analysis rely on the following technical ingredients.

Confidence regions based on minimax estimation via proximal causal inference. To handle the confounded offline dataset, we use the proxy variables from proximal causal inference (Lipsitch et al., 2010; Tchetgen et al., 2020; Miao et al., 2018a;b), which allows us to identify the value of each policy by a set of confounding bridge functions. These bridge functions only depend on observed variables and satisfy a set of backward conditional moment equations. We then estimate these bridge functions via minimax estimation (Dikkala et al., 2020; Chernozhukov et al., 2020; Uehara et al., 2021). More importantly, to handle the distributional shift, we propose a sequence of novel confidence regions for the bridge functions, which quantifies the uncertainty of minimax estimation based on finite data. This sequence of new confidence regions has not been considered in the previous works on off-policy evaluation (OPE) in POMDPs (Bennett and Kallus, 2021; Shi et al., 2021) as pessimism seems unnecessary in these works. Meanwhile, the confidence regions are constructed as a level set w.r.t. the loss functions of the minimax estimation for bridge functions. Such a construction contrasts with previous works on offline RL with confidence regions via maximum likelihood estimation (Uehara and Sun, 2021; Liu et al., 2022) or least square regression (Xie et al., 2021). Furthermore, we develop a novel theoretical analysis to show that any function in the confidence regions enjoys a fast statistical rate of convergence. Finally, leveraging the backwardly inductive nature of the bridge functions, our proposed confidence regions and analysis take the temporal structure into consideration, which might be of independent interest to the study on dynamic causal inference (Friston et al., 2003).

Pessimism principle for learning POMDPs. To learn the optimal policy in the face of distributional shift, we adopt the pessimism principle which is shown to be effective in offline RL in MDPs (Liu et al., 2020; Jin et al., 2021; Rashidinejad et al., 2021; Uehara and Sun, 2021; Xie et al., 2021; Yin and Wang, 2021; Zanette et al., 2021; Yin et al., 2022; Yan et al., 2022). Specifically, the newly proposed confidence regions, combined with the identification result based on proximal causal inference, allows us to construct a novel pessimistic estimator for the value of each target policy. From a theoretical perspective, the identification result and the backward induction property of the bridge functions provide a way of decomposing the suboptimality of the learned policy in terms of statistical errors of the bridge functions. When combined with the pessimism and the fast statistical rates enjoyed by any functions in the confidence regions, we show that our proposed P3O algorithm efficiently learns the optimal policy under only a partial coverage assumption of the confounded dataset. We highlight that our work firstly extend the pessimism principle to offline RL in POMDPs with confounded data.

1.2 Related Work

Our work is closely related to the bodies of literature on (i) reinforcement learning POMDPs, (ii) offline reinforcement learning (in MDPs), and (iii) OPE via causal inference. Compared to the

literature, our work simultaneously involve partial observability, confounded data, and offline policy optimization simultaneously, and thus involves the challenges faced by (i)–(iii). We summarize and contrast with most related existing works in Table 1. We defer the detailed discussion to Appendix B.

2 PRELIMINARIES

Notations. In the sequel, we use lower case letters (i.e., s, a, o, and τ) to represent dummy variables and upper case letters (i.e., S, A, O, and Γ) to represent random variables. We use the variables in the calligraphic font (i.e., $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{O}$) to represent the spaces of variables, and the blackboard bold font (i.e., $\mathbb{P}$ and $\mathbb{O}$) to represent probability kernels.

2.1 EPISODIC PARTIALLY OBSERVABLE MARKOV DECISION PROCESS

We consider an episodic, finite-horizon POMDP, specified by a tuple $(\mathcal{S}, \mathcal{O}, \mathcal{A}, H, \mu_1, \mathbb{P}, \mathbb{O}, R)$. Here we let $\mathcal{S}$, $\mathcal{A}$, and $\mathcal{O}$ denote the state, action, and observation spaces, respectively. The integer $H \in \mathbb{N}$ denotes the length of each episode. The distribution $\mu_1 \in \Delta(\mathcal{S})$ denotes the distribution of the initial state. The set $\mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}$ denotes the collection of state transition kernels where each kernel $\mathbb{P}_h(\cdot|s,a) : \mathcal{S} \times \mathcal{A} \mapsto \Delta(\mathcal{S})$ characterizes the distribution of the next state s_{h+1} given that the agent takes action $a_h = a \in \mathcal{A}$ at state $s_h = s \in \mathcal{S}$ and step $h \in [H]$. The set $\mathbb{O} = \{\mathbb{O}_h\}_{h=1}^H$ denotes the observation emission kernels where each kernel $\mathbb{O}_h(\cdot|s) : \mathcal{S} \mapsto \Delta(\mathcal{O})$ characterizes the distribution over observations given the current state $s \in \mathcal{S}$ at step $h \in [H]$. Finally, the set $R = \{R_h\}_{h=1}^H$ denotes the collection of reward functions where each function $R_h(s,a) : \mathcal{S} \times \mathcal{A} \mapsto [0,1]$ specifies the reward the agent receives when taking action $a \in \mathcal{A}$ at state $s \in \mathcal{S}$ and step $h \in [H]$.

Different from an MDP, in a POMDP, only the observation o, the action a, and the reward r are observable, while the state variable s is unobservable. In each episode, the environment first samples an initial state S_1 from $\mu_1(\cdot)$. At each step $h \in [H]$, the environment emits an observation O_h from $\mathbb{O}_h(\cdot|S_h)$. If an action A_h is taken, then the environment samples the next state S_{h+1} from $\mathbb{P}_h(\cdot|S_h, A_h)$ and assign a reward R_h given by $R_h(S_h, A_h)$. In our setting, we also let $O_0 \in \mathcal{O}$ denote the prior observation before step $h = 1$. We assume that O_0 is independent of other random variables in this episode given the first state S_1.

2.2 OFFLINE DATA GENERATION: CONFOUNDED DATASET

Now we describe the data generation process. Motivated by real-world examples such as precision medicine and autonomous driving discussed in Section 1, we assume that the offline data is generated by some behavior policy π^b which has access to the latent states. Specifically, we let $\pi^b = \{\pi_h^b\}_{h=1}^H$ denote a collection of policies such that $\pi_h^b(\cdot|s) : \mathcal{S} \mapsto \Delta(\mathcal{A})$ specifies the probability of taking action each $a \in \mathcal{A}$ at state s and step h. This behavior policy induces a set of probability distribution $\mathcal{P}^b = \{\mathcal{P}_h^b\}_{h=1}^H$ on the trajectories of the POMDP, where $\mathcal{P}_h^b$ is the distribution of the variables at step h when following the policy π^b. Formally, we assume that the offline data is denoted by $\mathbb{D} = \{(o_0^k, (o_1^k, a_1^k, r_1^k), \cdots, (o_H^k, a_H^k, r_H^k))\}_{k=1}^n$, where n is the number of trajectories, and for each $k \in [n]$, $(o_0^k, (o_1^k, a_1^k, r_1^k), \cdots, (o_H^k, a_H^k, r_H^k))$ is independently sampled from $\mathcal{P}^b$. We highlight that such an offline dataset is confounded since the latent state S_h, which is not stored in the dataset, simultaneously affects the control variables (i.e., action A_h) and the outcome variables (i.e., observation O_h and reward R_h). Such a setting is prohibitive for existing offline RL algorithms for MDPs as directly applying them will nevertheless incur a confounding bias that is not negligible.

2.3 LEARNING OBJECTIVE

The goal of offline RL is to learn an optimal policy from the offline dataset which maximizes the expected cumulative reward. For POMDPs, the learned policy can only depend on the observable information mentioned in Section 2.1. To formally define the set policies of interest, we first define the space of observable history as $\mathcal{H} = \{\mathcal{H}_h\}_{h=0}^{H-1}$, where each element $\tau_h \in \mathcal{H}_h$ is a (partial) trajectory such that $\tau_h \subseteq \{(o_1, a_1), \cdots, (o_h, a_h)\}$. We use Γ_h to denote the corresponding random variable.

We denote by $\Pi(\mathcal{H})$ the class of policies which make decisions based on the current observation $o_h \in \mathcal{O}$ and the history information $\tau_{h-1} \in \mathcal{H}_{h-1}$. That means, a policy $\pi = \{\pi_h\}_{h=1}^H \in \Pi(\mathcal{H})$

satisfies $\pi_h(\cdot|o,\tau) : \mathcal{O}\times\mathcal{H}_{h-1} \mapsto \Delta(\mathcal{A})$. The choice of $\mathcal{H}$ induces the policy set $\Pi(\mathcal{H})$ by specifying the input of the policies. We now introduce three examples of $\mathcal{H}$ and the corresponding $\Pi(\mathcal{H})$.

Example 2.1 (Reactive policy (Azizzadenesheli et al., 2018))**.** *The policy only depends on the current observation O_h. Formally, we have $\mathcal{H}_{h-1} = \{\varnothing\}$ and therefore $\tau_{h-1} = \varnothing$ for each $h \in [H]$.*

Example 2.2 (Finite-history policy (Efroni et al., 2022))**.** *The policy depends on the current observation and the history of length at most k. Formally, we have $\mathcal{H}_{h-1} = (\mathcal{O}\times\mathcal{A})^{\otimes \min\{k,h-1\}}$ and $\tau_{h-1} = ((o_l, a_l), \cdots, (o_{h-1}, a_{h-1}))$ for some $k \in \mathbb{N}$, where the index $l = \max\{1, h-k\}$.*

Example 2.3 (Full-history policy (Liu et al., 2022))**.** *The policy depends on the current observation and the full history. Formally, $\mathcal{H}_{h-1} = (\mathcal{O}\times\mathcal{A})^{\otimes(h-1)}$, and $\tau_{h-1} = ((o_1, a_1), \cdots, (o_{h-1}, a_{h-1}))$.*

We illustrate these examples with causal graphs and a more detailed discussion in Appendix C.1. Now given a policy $\pi \in \Pi(\mathcal{H})$, we denote by $J(\pi)$ the value of π that characterizes the expected cumulative rewards the agent receives by following π. Formally, $J(\pi)$ is defined as

$$J(\pi) := \mathbb{E}_\pi\left[\sum_{h=1}^{H}\gamma^{h-1}R_h(S_h, A_h)\right], \tag{2.1}$$

where $\gamma \in (0,1]$ is the discount factor, $\mathbb{E}_\pi$ denotes the expectation w.r.t. $\mathcal{P}^\pi = \{\mathcal{P}_h^\pi\}_{h=1}^H$ which is the distribution of the trajectories induced by π. We define the suboptimality gap of any policy $\widehat{\pi}$ as

$$\text{SubOpt}(\widehat{\pi}) := J(\pi^\star) - J(\widehat{\pi}), \quad \text{where} \quad \pi^\star \in \operatorname*{arg\,max}_{\pi\in\Pi(\mathcal{H})} J(\pi). \tag{2.2}$$

Here $\pi^\star$ is the optimal policy within $\Pi(\mathcal{H})$. Our goal is to find some policy $\widehat{\pi} \in \Pi(\mathcal{H})$ that minimizes the suboptimality gap in (2.2) based on the offline dataset $\mathbb{D}$.

3 Algorithm: Proxy Variable Pessimistic Policy Optimization

The offline RL problem introduced in Section 2 for POMDPs suffers from three coupled challenges — (i) the confounding bias, (ii) distributional shift, and (iii) large observation spaces. Among these, the confounding bias is caused by the fact that the latent state S_h simultaneous affects the action variable (i.e., A_h) and the outcome variables (i.e., O_h and R_h). The distributional shift exists between the trajectories induced by the behavior policy π^b and the family of target policies $\Pi(\mathcal{H})$. The challenge of large observation spaces necessitates the adoption of function approximators. In the sequel, we introduce an algorithm that addresses all three challenges simultaneously.

Offline RL for POMDPs is known to be intractable in the worst case (Krishnamurthy et al., 2016). So we first identify a benign class of POMDPs where the causal structure involving latent states can be captured by only the observable variables available in the dataset $\mathbb{D}$. For such a class of POMDPs, by leveraging tools from proximal causal inference (Lipsitch et al., 2010; Tchetgen et al., 2020), we then seek to identify the value $J(\pi)$ of the policy $\pi \in \Pi(\mathcal{H})$ via some confounding bridge functions $\mathbf{b}^\pi$ (Assumption 3.2) which only depend on the observable variables and thus can be estimated using $\mathbb{D}$ (Theorem 3.3). Identification via proximal causal inference will be discussed in Section 3.1.

In addition, to estimate these confounding bridge functions, we utilize the the fact that these functions satisfy a sequence of conditional moment equations which resembles the Bellman equations in classical MDPs (Bellman and Kalaba, 1965). Then we adopt the idea of minimax estimation (Dikkala et al., 2020; Kallus et al., 2021; Uehara et al., 2021; Duan et al., 2021) which formulates the bridge functions as the solution to a series of minimax optimization problems in (3.11). Additionally, the loss function in minimax estimation readily incorporates function approximators and thus addresses the challenge of large observation spaces.

To further handle the distributional shift, we extend the pessimism principle (Liu et al., 2020; Jin et al., 2021; Rashidinejad et al., 2021; Uehara and Sun, 2021; Xie et al., 2021; Yin and Wang, 2021; Zanette et al., 2021) to POMDPs with the help of the confounding bridge functions. In specific, based on the confounded dataset, we first construct a novel confidence region $\text{CR}^\pi(\xi)$ for $\mathbf{b}^\pi$ based on level sets with respect to the loss functions of the minimax estimation (See (3.12) for details). Our algorithm, Proxy variable Pessimistic Policy Optimization (`P3O`), outputs the policy that maximizes pessimistic estimates the values of the policies within $\Pi(\mathcal{H})$. The details of `P3O` is summarized by Algorithm 1 in Section 3.3.

3.1 POLICY VALUE IDENTIFICATION VIA PROXIMAL CAUSAL INFERENCE

To handle the confounded dataset $\mathbb{D}$, we first identify the policy value $J(\pi)$ for each $\pi \in \Pi(\mathcal{H})$ using the idea of proxy variables. Following the notions of proximal causal inference (Tennenholtz et al., 2020; Lipsitch et al., 2010), we assume that there exists negative control actions $\{Z_h\}_{h=1}^H$ and negative control outcomes $\{W_h\}_{h=1}^H$ satisfying the following independence assumption.

Assumption 3.1 (Negative control)**.** *We assume there exist negative control variables $\{W_h\}_{h=1}^H$ and $\{Z_h\}_{h=1}^H$ measurable with respect to the observed trajectories, such that under $\mathcal{P}^b$, it holds that*

$$Z_h \perp O_h, R_h, W_h, W_{h+1} \mid A_h, S_h, \Gamma_{h-1}, \quad W_h \perp A_h, \Gamma_{h-1}, S_{h-1} \mid S_h. \tag{3.1}$$

We explain in detail the existence of such negative control variables for all the three different choices of history $\mathcal{H}$ in Examples 2.1, 2.2, and 2.3 respectively in Appendix C.1. Besides Assumption 3.1, our identification of policy value also relies on the notion of confounding bridge functions (Kallus et al., 2021; Shi et al., 2021), for which we make the following assumption.

Assumption 3.2 (Confounding bridge functions)**.** *For any history-dependent policy $\pi \in \Pi(\mathcal{H})$, we assume the existence of the value bridge functions $\{b_h^\pi : \mathcal{A} \times \mathcal{W} \mapsto \mathbb{R}\}_{h=1}^H$ and the weight bridge functions $\{q_h^\pi : \mathcal{A} \times \mathcal{Z} \mapsto \mathbb{R}\}_{h=1}^H$ which are defined as the solution to the following equations almost surely with respect to the measure $\mathcal{P}^b$:*

$$\begin{aligned}\mathbb{E}_{\pi^b}\left[b_h^\pi(A_h, W_h)|A_h, Z_h\right] = \mathbb{E}_{\pi^b}\Big[&R_h \pi_h(A_h|O_h, \Gamma_{h-1}) \\ &+ \gamma \sum_{a'} b_{h+1}^\pi(a', W_{h+1})\pi_h(A_h|O_h, \Gamma_{h-1})\Big| A_h, Z_h\Big],\end{aligned} \tag{3.2}$$

$$\mathbb{E}_{\pi^b}\left[q_h^\pi(A_h, Z_h)|A_h, S_h, \Gamma_{h-1}\right] = \frac{\mu_h(S_h, \Gamma_{h-1})}{\pi_h^b(A_h|S_h)}. \tag{3.3}$$

Here b_{H+1}^π is a zero function and $\mu_h(S_h, \Gamma_{h-1})$ in (3.3) is defined as the importance sampling ratio $\mu_h(S_h, \Gamma_{h-1}) \coloneqq \mathcal{P}_h^\pi(S_h, \Gamma_{h-1})/\mathcal{P}_h^b(S_h, \Gamma_{h-1})$.

We use "confounding bridge function" and "bridge function" interchangeably throughout the paper. We remark that in the proximal causal inference literature, the existence of such bridge functions bears more generality than assuming certain complicated completeness conditions (Cui et al., 2020), as discussed by Kallus et al. (2021). The existence of such bridge functions is justified by conditions on the the rank of certain conditional probabilities or singular values of certain conditional expectation linear operators. See Appendix C.2 for examples of Assmuption 3.2 under previous choice of $\mathcal{H}$.

Now given Assumption 3.1 and 3.2 on the existence of proxy variables and bridge functions, we are ready to present the main identification result. It represents the true policy value $J(\pi)$ via the value bridge functions, as is concluded in the following theorem.

Theorem 3.3 (Identification of policy value)**.** *Under Assumption 3.1 and 3.2, for any history-dependent policy $\pi \in \Pi(\mathcal{H})$, it holds that*

$$J(\pi) = F(\mathbf{b}^\pi), \quad \text{where} \quad F(\mathbf{b}^\pi) \coloneqq \mathbb{E}_{\pi^b}\left[\sum_{a \in \mathcal{A}} b_1^\pi(a, W_1)\right]. \tag{3.4}$$

See Appendix E for a detailed proof. Note that although we have assumed the existence of both the value bridge functions in (3.2) and the weight bridge functions in (3.3), Theorem 3.3 represents $J(\pi)$ using only the value bridge functions. In (3.2) all the random variables involved are observed by the learner and distributed according to the data distribution $\mathcal{P}^b$, which means that the value bridge functions can be estimated from data. This overcomes the confounding issue.

3.2 POLICY EVALUATION VIA MINIMAX ESTIMATION WITH UNCERTAINTY QUANTIFICATION

According to Theorem 3.3 and Assumption 3.2, to estimate the value $J(\pi)$ of $\pi \in \Pi(\mathcal{H})$, it suffices to estimate the value bridge functions $\{b_h^\pi\}_{h=1}^H$ by solving (3.2), which is a conditional moment equation. To this end, we adopt the method of minimax estimation (Dikkala et al., 2020; Uehara et al., 2021; Duan et al., 2021). Furthermore, in order to handle the distributional shift between behavior

policy and target policies, we construct a sequence of confidence regions for $\{b_h^\pi\}_{h=1}^H$ based on minimax estimation, which allows us to apply the pessimism principle by finding the most pessimistic estimates within the confidence regions.

Specifically, minimax estimation involves two function classes $\mathbb{B} \subseteq \{b : \mathcal{A} \times \mathcal{W} \mapsto \mathbb{R}\}$ and $\mathbb{G} \subseteq \{g : \mathcal{A} \times \mathcal{Z} \mapsto \mathbb{R}\}$, interpreted as the primal and dual function classes, respectively. Theoretical assumptions on $\mathbb{B}$ and $\mathbb{G}$ are presented in Section 4. In order to find functions that satisfy (3.2), it suffices to find $\mathbf{b} = (b_1, \cdots, b_H) \in \mathbb{B}^{\otimes H}$ such that the following conditional moment

$$\ell_h^\pi(b_h, b_{h+1})(A_h, Z_h) \coloneqq \mathbb{E}_{\pi^b}\Big[b_h(A_h, W_h) - R_h\pi_h(A_h|O_h, \Gamma_{h-1}) - \gamma \sum_{a' \in \mathcal{A}} b_{h+1}(a', W_{h+1})\pi_h(A_h|O_h, \Gamma_{h-1})\Big| A_h, Z_h\Big] \tag{3.5}$$

is equal to zero almost surely for all $h \in [H]$, where b_{H+1} is a zero function. Intuitively, the quantity (3.5) can be interpreted as the "Bellman residual" of the value bridge functions $\mathbf{b}$. Notice that (3.5) being zero almost surely for all $h \in [H]$ is actually a conditional moment equation, which is equivalent to finding $\mathbf{b} \in \mathbb{B}^{\otimes H}$ such that the following residual mean squared error (RMSE) is minimized for all h:

$$\mathcal{L}_h^\pi(b_h, b_{h+1}) \coloneqq \mathbb{E}_{\pi^b}\big[\big(\ell_h^\pi(b_h, b_{h+1})(A_h, Z_h)\big)^2\big]. \tag{3.6}$$

It might seem tempting to directly minimize the empirical version of (3.6). However, this is not viable as one would obtain a biased estimator due to an additional variance term. The reason is that the quantity defined by (3.5) is a conditional expectation and therefore RMSE defined by (3.6) cannot be directly unbiasedly estimated from data (Farahmand et al., 2016). In the sequel, we adopt the technique of minimax estimation to circumvent this issue. In particular, we first use Fenchel duality to write (3.6) as

$$\mathcal{L}_h^\pi(b_h, b_{h+1}) = 4\lambda\mathbb{E}_{\pi^b}\left[\max_{g \in \mathbb{G}} \ell_h^\pi(b_h, b_{h+1})(A_h, Z_h) \cdot g(A_h, Z_h) - \lambda g(A_h, Z_h)^2\right],\ \lambda > 0, \tag{3.7}$$

which holds when the dual function class $\mathbb{G}$ is expressive enough such that $\ell_h^\pi(b_h, b_{h+1})/2\lambda \in \mathbb{G}$. Then thanks to the interchangeability principle (Rockafellar and Wets, 2009; Dai et al., 2017; Shapiro et al., 2021), we can interchange the order of maximization and expectation and derive that

$$\mathcal{L}_h^\pi(b_h, b_{h+1}) = 4\lambda \max_{g \in \mathbb{G}} \mathbb{E}_{\pi^b}\left[\ell_h^\pi(b_h, b_{h+1})(A_h, Z_h) \cdot g(A_h, Z_h) - \lambda g(A_h, Z_h)^2\right]. \tag{3.8}$$

The core idea of minimax estimation is to minimize the empirical version of (3.8) instead of (3.6), and the benefit of doing so is a fast statistical rate of $\widetilde{\mathcal{O}}(n^{-1/2})$ (Dikkala et al., 2020; Uehara et al., 2021), as we can see in the sequel. For simplicity, in the following, we define $\Phi_{\pi,h}^\lambda : \mathbb{B} \times \mathbb{B} \times \mathbb{G} \mapsto \mathbb{R}$ with parameter $\lambda > 0$ as

$$\Phi_{\pi,h}^\lambda(b_h, b_{h+1}; g) \coloneqq \mathbb{E}_{\pi^b}\big[\ell_h^\pi(b_h, b_{h+1})(A_h, Z_h) \cdot g(A_h, Z_h) - \lambda g(A_h, Z_h)^2\big], \tag{3.9}$$

and we denote by $\widehat{\Phi}_{\pi,h}^\lambda : \mathbb{B} \times \mathbb{B} \times \mathbb{G} \mapsto \mathbb{R}$ the empirical version of $\Phi_{\pi,h}^\lambda$, i.e.,

$$\widehat{\Phi}_{\pi,h}^\lambda(b_h, b_{h+1}; g) \coloneqq \widehat{\mathbb{E}}_{\pi^b}\Big[\Big(b_h(A_h, W_h) - R_h\pi_h(A_h|O_h, \Gamma_{h-1}) - \gamma \sum_{a' \in \mathcal{A}} b_{h+1}(a', W_{h+1})\pi_h(A_h|O_h, \Gamma_{h-1})\Big) \cdot g(A_h, Z_h) - \lambda g(A_h, Z_h)^2\Big], \tag{3.10}$$

where $\widehat{\mathbb{E}}_{\pi^b}$ denotes the empirical version of $\mathbb{E}_{\pi^b}$ based on dataset $\mathbb{D}$ described in Section 2.2.

Furthermore, note that the value bridge functions $(b_1^\pi, \cdots, b_h^\pi)$ admit a sequential dependence structure. To handle such dependency, for any $\pi \in \Pi(\mathcal{H})$, $h \in [H]$, and $b_{h+1} \in \mathbb{B}$, we first define the minimax estimator $\widehat{b}_h(b_{h+1})$ as

$$\widehat{b}_h(b_{h+1}) \coloneqq \arg\min_{b \in \mathbb{B}}\Big\{\max_{g \in \mathbb{G}} \widehat{\Phi}_{\pi,h}^\lambda(b_h, b_{h+1}; g)\Big\}. \tag{3.11}$$

Based on (3.11), we propose a confidence region for $\mathbf{b}^\pi \coloneqq (b_1^\pi, \cdots, b_H^\pi) \in \mathbb{B}^{\otimes H}$ as

$$\mathrm{CR}^\pi(\xi) \coloneqq \left\{\mathbf{b} \in \mathbb{B}^{\otimes H}\,\middle|\, \max_{g \in \mathbb{G}} \widehat{\Phi}_{\pi,h}^\lambda(b_h, b_{h+1}; g) - \max_{g \in \mathbb{G}} \widehat{\Phi}_{\pi,h}^\lambda(\widehat{b}_h(b_{h+1}), b_{h+1}; g) \le \xi, \forall h\right\}. \tag{3.12}$$

Algorithm 1 Proxy variable Pessimistic Policy Optimization (P3O)

1: **Input**: confidence parameter $\xi > 0$, regularization parameter $\lambda > 0$, dataset $\mathbb{D}$, classes $\mathbb{B}$ and $\mathbb{G}$.
2: Construct minimax estimator confidence region $\mathrm{CR}^{\pi}(\xi)$ for each $\pi \in \Pi(\mathcal{H})$ by (3.12).
3: **Policy evaluation:** pessimistically estimate $\widehat{J}_{\mathrm{Pess}}(\pi)$ for each $\pi \in \Pi(\mathcal{H})$ by (3.13).
4: **Policy optimization:** set $\widehat{\pi}$ by (3.14).
5: **Output**: $\widehat{\pi} = \{\widehat{\pi}_h\}_{h=1}^{H}$.

From the above definition[1], one can see that $\mathrm{CR}^{\pi}(\xi)$ is actually a coupled sequence of H confidence regions, where each single confidence region aims to cover a function b_h^{π}. For notational simplicity, we use a single notation $\mathrm{CR}^{\pi}(\xi)$ to denote all the H confidence regions. Intuitively, the confidence region $\mathrm{CR}^{\pi}(\xi)$ contains all $\mathbf{b} \in \mathbb{B}^{\otimes H}$ whose RMSE does not exceed that of $(\widehat{b}_h(b_{h+1}),\ b_{h+1})$ by too much at each $h \in [H]$. The confidence region takes the sequential dependence of confounding bridge functions into consideration in the sense that each $\mathbf{b} \in \mathrm{CR}^{\pi}(\xi)$ is restricted through the minimax estimation loss between continuous steps. As we show in Section D, with high probability, the confidence region $\mathrm{CR}^{\pi}(\xi)$ contains the true bridge value functions $\mathbf{b}^{\pi}$. More importantly, every $\mathbf{b} \in \mathrm{CR}^{\pi}(\xi)$ enjoys a fast statistical rate of $\widetilde{\mathcal{O}}(n^{-1/2})$.

Now combining the confidence region (3.12) and the identification formula (3.4), for any policy $\pi \in \Pi(\mathcal{H})$, we adopt an pessimistic estimate of the value of $J(\pi)$ as

$$\widehat{J}_{\mathrm{Pess}}(\pi) := \min_{\mathbf{b} \in \mathrm{CR}^{\pi}(\xi)} \widehat{F}(\mathbf{b}), \quad \text{where} \quad \widehat{F}(\mathbf{b}) := \widehat{\mathbb{E}}_{\pi^b}\left[\sum_{a \in \mathcal{A}} b_1(a, W_1)\right]. \tag{3.13}$$

3.3 Policy Optimization

Given the pessimistic value estimate (3.13), P3O chooses $\widehat{\pi}$ which maximizes $\widehat{J}_{\mathrm{Pess}}(\pi)$, that is,

$$\widehat{\pi} := \operatorname*{arg\,max}_{\pi \in \Pi(\mathcal{H})} \widehat{J}_{\mathrm{Pess}}(\pi). \tag{3.14}$$

We summarize the entire P3O algorithm in Algorithm 1. In Section 4, we show that under some mild assumptions on the function classes $\mathbb{B}$ and $\mathbb{G}$ and under only a partial coverage assumption on the dataset $\mathbb{D}$, the suboptimality (2.2) of Algorithm 1 decays at the fast statistical rate of $\widetilde{\mathcal{O}}(n^{-1/2})$, where $\widetilde{\mathcal{O}}(\cdot)$ omits H and factors that characterize the complexity of the function classes.

4 Theoretical Results

In this section, we present our theoretical results. For ease of presentation, we first assume that both the primal function class $\mathbb{B}$ and the policy class $\Pi(\mathcal{H})$ are finite sets with cardinality $|\mathbb{B}|$ and $|\Pi(\mathcal{H})|$, respectively. But we allow the dual function class $\mathbb{G}$ to be an infinite set. Our results can be easily extended to infinite $\mathbb{B}$ and $\Pi(\mathcal{H})$ using the notion of covering numbers (Wainwright, 2019), which we demonstrate with linear function approximation in Section H.1.

We first introduce some necessary assumptions for efficient learning of the optimal policy. To begin with, the following Assumption 4.1 ensures that the offline data generated by π^b has a good coverage over $\pi^\star$. The problem would become intractable without such an assumption (Chen and Jiang, 2019).

Assumption 4.1 (Partial coverage)**.** *We assume that the concentrability coefficient for the optimal policy $\pi^\star$, defined as $C^{\pi^\star} := \max_{h \in [H]} \mathbb{E}_{\pi^b}\left[(q_h^{\pi^\star}(A_h, Z_h))^2\right]$, satisfies that $C^{\pi^\star} < +\infty$.*

Very importantly, Assumption 4.1 only assumes the partial coverage, i.e., the optimal policy $\pi^\star$ is well covered by π^b (Jin et al., 2021; Uehara and Sun, 2021), which is significantly weaker than the uniform coverage, i.e., the entire policy class $\Pi(\mathcal{H})$ is covered by π^b (Chen and Jiang, 2019) in the sense that $\max_{\pi \in \Pi(\mathcal{H})} C^{\pi} < +\infty$. See Appendix B for more about partial coverage in POMDP.

The next assumption is on the functions classes $\mathbb{B}$ and $\mathbb{G}$. We require that $\mathbb{B}$ and $\mathbb{G}$ are uniformly bounded, and that $\mathbb{G}$ is symmetric, star-shaped, and has bounded localized Rademacher complexity.

[1] We refer the readers to Appendix B for a detailed comparison of the minimax-typed loss (and confidence region) with the least-square-typed loss (and confidence region) used by Xie et al. (2021).

Assumption 4.2 (Function classes $\mathbb{B}$ and $\mathbb{G}$)**.** *We assume the classes $\mathbb{B}$ and $\mathbb{G}$ satisfy that: i) There exist $M_{\mathbb{B}}, M_{\mathbb{G}} < +\infty$ such that $\mathbb{B}$, $\mathbb{G}$ are bounded by $\sup_{b\in\mathbb{B}}\sup_{w\in\mathcal{W}}|\sum_{a\in\mathcal{A}} b(a,w)| \leq M_{\mathbb{B}}$[2], $\sup_{g\in\mathbb{G}}\sup_{(a,z)\in\mathcal{A}\times\mathcal{Z}}|g(a,z)| \leq M_{\mathbb{G}}$; ii) $\mathbb{G}$ is star-shaped, i.e., for any $g \in \mathbb{G}$ and $\lambda \in [0,1]$, it holds that $\lambda g \in \mathbb{G}$; iii) $\mathbb{G}$ is symmetric, i.e., for any $g \in \mathbb{G}$, it holds that $-g \in \mathbb{G}$; iv) For any step $h \in [H]$, $\mathbb{G}$ has bounded critical radius $\alpha_{\mathbb{G},h,n}$ which solves inequality $\mathcal{R}_n(\mathbb{G};\alpha) \leq \alpha^2/M_{\mathbb{G}}$, where $\mathcal{R}_n(\mathbb{G},\alpha)$ is the localized population Rademacher complexity of $\mathbb{G}$ under the distribution of (A_h, Z_h) induced by π^b, that is,*

$$\mathcal{R}_n(\mathbb{G},\alpha) = \mathbb{E}_{\pi^b,\epsilon_i}\left[\sup_{g\in\mathbb{G}:\|g\|_2\leq\alpha}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i g(A_h,Z_h)\right|\right],$$

with $\|g\|_2$ defined as $(\mathbb{E}_{\pi^b}[g^2(A_h,Z_h)])^{1/2}$, random variables $\{\epsilon_i\}_{i=1}^n$ independent of (A_h,Z_h) and independently uniformly distributed on $\{+1,-1\}$. Also, we denote $\alpha_{\mathbb{G},n} \coloneqq \max_{h\in[H]}\alpha_{\mathbb{G},h,n}$.

Finally, to ensure that the minimax estimation in Section 3.2 learns the value bridge functions unbiasedly, we make the following completeness and realizability assumptions on the function classes $\mathbb{B}$ and $\mathbb{G}$ which are standard in the literature (Dikkala et al., 2020; Xie et al., 2021; Shi et al., 2021).

Assumption 4.3 (Completeness and realizability)**.** *We assume that, i) completeness: for any $h \in [H]$, any $\pi \in \Pi(\mathcal{H})$, and any $b_h, b_{h+1} \in \mathbb{B}$, it holds that $\frac{1}{2\lambda}\ell_h^\pi(b_h,b_{h+1}) \in \mathbb{G}$ where ℓ_h^π is defined in (3.5); ii) realizability: for any $h \in [H]$, any $\pi \in \Pi(\mathcal{H})$, and any $b_{h+1} \in \mathbb{B}$, there exists $b^\star \in \mathbb{B}$ such that $\mathcal{L}_h^\pi(b^\star, b_{h+1}) \leq \epsilon_{\mathbb{B}}$ for some $\epsilon_{\mathbb{B}} < +\infty$, i.e., we assume that*

$$0 \leq \epsilon_{\mathbb{B}} \coloneqq \max_{h\in[H],\pi\in\Pi(\mathcal{H}),b_{h+1}\in\mathbb{B}}\ \min_{b_h\in\mathbb{B}}\ \mathbb{E}_{\pi^b}\left[\ell_h^\pi(b_h,b_{h+1})(A_h,Z_h)^2\right] < +\infty.$$

Here the completeness assumption means the dual function class $\mathbb{G}$ is rich enough which guarantees the equivalence between $\mathcal{L}_h^\pi(\cdot,\cdot)$ and $\max_{g\in\mathbb{G}}\Phi_{\pi,h}^\lambda(\cdot,\cdot;g)$. The realizability assumption means that the primal function class $\mathbb{B}$ is rich enough such that (3.2) always admits an (approximate) solution.

With these technical assumptions, we can establish our main theoretical results in the following theorem, which gives an upper bound of the suboptimality (2.2) of the policy $\widehat{\pi}$ output by Algorithm 1.

Theorem 4.4 (Suboptimality)**.** *Under Assumptions 3.1, 3.2, 4.1, 4.2, and 4.3, by setting the regularization parameter λ and the confidence parameter ξ as $\lambda = 1$ and*

$$\xi = C_1 \cdot M_{\mathbb{B}}^2 M_{\mathbb{G}}^2 \cdot \log(|\mathbb{B}||\Pi(\mathcal{H})|H/\zeta)/n,$$

then probability at least $1-3\delta$, it holds that

$$\mathrm{SubOpt}(\widehat{\pi}) \leq C_1'\sqrt{C^{\pi^\star}}H \cdot M_{\mathbb{B}}M_{\mathbb{G}} \cdot \sqrt{\log(|\mathbb{B}||\Pi(\mathcal{H})|H/\zeta)/n} + C_1'\sqrt{C^{\pi^\star}M_{\mathbb{G}}}H\epsilon_{\mathbb{B}}^{1/4},$$

where $\zeta = \min\{\delta, 4c_1\exp(-c_2 n\alpha_{\mathbb{G},n}^2)\}$. Here $C^{\pi^\star}$, $\alpha_{\mathbb{G},n}$, and $\epsilon_{\mathbb{B}}$ are defined in Assumption 4.1, 4.2, and 4.3, respectively. And C_1, C_1', c_1, and c_2 are some problem-independent universal constants.

We introduce all the key technical lemmas and sketch the proof of Theorem 4.4 in Section D. We refer to Appendix G for a detailed proof. When it holds that $\alpha_{\mathbb{G},n} = \widetilde{\mathcal{O}}(n^{-1/2})$ and $\epsilon_{\mathbb{B}} = 0$, Theorem 4.4 implies that $\mathrm{SubOpt}(\widehat{\pi}) \leq \widetilde{\mathcal{O}}(n^{-1/2})$, which corresponds to a "fast statistical rate" for minimax estimation (Uehara et al., 2021). The derivation of such a fast rate relies on a novel analysis for the risk of functions in the confidence region, which is shown by Lemma D.3 in Section D. Meanwhile, for many choices of the dual function class $\mathbb{G}$, it holds that $\alpha_{\mathbb{G},n}$ scales with $\sqrt{\log\mathcal{N}_{\mathbb{G}}}$ where $\mathcal{N}_{\mathbb{G}}$ denotes the complexity measure of the class $\mathbb{G}$. In such cases, the suboptimality also scales with $\sqrt{\log\mathcal{N}_{\mathbb{G}}}$, without explicit dependence on the cardinality of the spaces $\mathcal{S}$, $\mathcal{A}$, or $\mathcal{O}$. Finally, we highlight that, thanks to the principle of pessimism, the suboptimality of `P3O` depends only on the partial coverage concentrability coefficient $C^{\pi^\star}$, which can be significantly smaller than the uniform coverage concentrability coefficient $\sup_{\pi\in\Pi(\mathcal{H})} C^\pi$. In conclusion, when $\alpha_{\mathbb{G},n} = \widetilde{\mathcal{O}}(\sqrt{\log\mathcal{N}_{\mathbb{G}}/n})$ and $\epsilon_{\mathbb{B}} = 0$, the `P3O` algorithm enjoys a $\widetilde{\mathcal{O}}(H\sqrt{C^{\pi^\star}\log\mathcal{N}_{\mathbb{G}}/n})$ suboptimality.

Linear function approximation. Theorem 4.4 can be readily extended to the case of linear function approximation (LFA) with infinite-cardinality $\mathbb{B}$ and $\Pi(\mathcal{H})$, which yields an $\widetilde{\mathcal{O}}(\sqrt{H^3 d/n})$ suboptimality. Due to space limit, we defer the detailed setup and main results of LFA to Appendix H.1.

[2]The constant $\mathrm{M}_{\mathbb{B}}$ might seem to be proportional to $|\mathcal{A}|$ due to the summation $\sum_{a\in\mathcal{A}} b(a,w)$, but it is not. The reason is that the definition of the true confounding bridge function in (3.2) involves a product with $\pi_h(\cdot|O_h,\Gamma_{h-1})$ which is a distribution over $\mathcal{A}$. Thus the summation over $\mathcal{A}$ is essentially an average over $\mathcal{A}$.

REFERENCES

ANTOS, A., MUNOS, R. and SZEPESVÁRI, C. (2007). Fitted q-iteration in continuous action-space mdps .

AZIZZADENESHELI, K., YUE, Y. and ANANDKUMAR, A. (2018). Policy gradient in partially observable environments: Approximation and convergence. *arXiv preprint arXiv:1810.07900* .

BELLMAN, R. and KALABA, R. E. (1965). *Dynamic programming and modern control theory*, vol. 81. Citeseer.

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

BENNETT, A., KALLUS, N., LI, L. and MOUSAVI, A. (2021). Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

BUCKMAN, J., GELADA, C. and BELLEMARE, M. G. (2020). The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799* .

BURAGO, D., DE ROUGEMONT, M. and SLISSENKO, A. (1996). On the complexity of partially observed Markov decision processes. *Theoretical Computer Science* **157** 161–183.

CHAKRABORTY, B. and MOODIE, E. (2013). Statistical methods for dynamic treatment regimes. *Springer-Verlag. doi* **10** 978–1.

CHAKRABORTY, B. and MURPHY, S. A. (2014). Dynamic treatment regimes. *Annual review of statistics and its application* **1** 447–464.

CHEN, J. and JIANG, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*. PMLR.

CHERNOZHUKOV, V., NEWEY, W., SINGH, R. and SYRGKANIS, V. (2020). Adversarial estimation of riesz representers. *arXiv preprint arXiv:2101.00009* .

CUI, Y., PU, H., SHI, X., MIAO, W. and TCHETGEN, E. T. (2020). Semiparametric proximal causal inference. *arXiv preprint arXiv:2011.08411* .

DAI, B., HE, N., PAN, Y., BOOTS, B. and SONG, L. (2017). Learning from conditional distributions via dual embeddings. In *Artificial Intelligence and Statistics*. PMLR.

DIKKALA, N., LEWIS, G., MACKEY, L. and SYRGKANIS, V. (2020). Minimax estimation of conditional moment models. *Advances in Neural Information Processing Systems* **33** 12248–12262.

DUAN, Y., JIN, C. and LI, Z. (2021). Risk bounds and rademacher complexity in batch reinforcement learning. In *International Conference on Machine Learning*. PMLR.

DULAC-ARNOLD, G., LEVINE, N., MANKOWITZ, D. J., LI, J., PADURARU, C., GOWAL, S. and HESTER, T. (2021). Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning* **110** 2419–2468.

EFRONI, Y., JIN, C., KRISHNAMURTHY, A. and MIRYOOSEFI, S. (2022). Provable reinforcement learning with a short-term memory. *arXiv preprint arXiv:2202.03983* .

FARAHMAND, A.-M., GHAVAMZADEH, M., SZEPESVÁRI, C. and MANNOR, S. (2016). Regularized policy iteration with nonparametric function spaces. *The Journal of Machine Learning Research* **17** 4809–4874.

FOSTER, D. J. and SYRGKANIS, V. (2019). Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036* .

FRISTON, K. J., HARRISON, L. and PENNY, W. (2003). Dynamic causal modelling. *Neuroimage* **19** 1273–1302.

GOLDSMITH, J. and MUNDHENK, M. (1998). Complexity issues in Markov decision processes. In *Computational Complexity Conference*.

GOTTESMAN, O., JOHANSSON, F., KOMOROWSKI, M., FAISAL, A., SONTAG, D., DOSHI-VELEZ, F. and CELI, L. A. (2019). Guidelines for reinforcement learning in healthcare. *Nature medicine* **25** 16–18.

GUO, Z. D., DOROUDI, S. and BRUNSKILL, E. (2016). A pac rl algorithm for episodic pomdps. In *Artificial Intelligence and Statistics*. PMLR.

JAFARNIA-JAHROMI, M., JAIN, R. and NAYYAR, A. (2021). Online learning for unknown partially observable mdps. *arXiv preprint arXiv:2102.12661* .

JANNER, M., LI, Q. and LEVINE, S. (2021). Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems* **34**.

JIN, C., KAKADE, S., KRISHNAMURTHY, A. and LIU, Q. (2020). Sample-efficient reinforcement learning of undercomplete pomdps. *Advances in Neural Information Processing Systems* **33** 18530–18539.

JIN, Y., YANG, Z. and WANG, Z. (2021). Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*. PMLR.

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

KALLUS, N. and ZHOU, A. (2020). Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Advances in Neural Information Processing Systems* **33** 22293–22304.

KALLUS, N. and ZHOU, A. (2021). Minimax-optimal policy learning under unobserved confounding. *Management Science* **67** 2870–2890.

KIDAMBI, R., RAJESWARAN, A., NETRAPALLI, P. and JOACHIMS, T. (2020). Morel: Model-based offline reinforcement learning. *arXiv preprint arXiv:2005.05951* .

KRISHNAMURTHY, A., AGARWAL, A. and LANGFORD, J. (2016). Pac reinforcement learning with rich observations. *Advances in Neural Information Processing Systems* **29**.

KUMAR, A., ZHOU, A., TUCKER, G. and LEVINE, S. (2020). Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems* **33** 1179–1191.

LEVINE, S., KUMAR, A., TUCKER, G. and FU, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643* .

LIPSITCH, M., TCHETGEN, E. T. and COHEN, T. (2010). Negative controls: a tool for detecting confounding and bias in observational studies. *Epidemiology (Cambridge, Mass.)* **21** 383.

LIU, Q., CHUNG, A., SZEPESVÁRI, C. and JIN, C. (2022). When is partially observable reinforcement learning not scary? *arXiv preprint arXiv:2204.08967* .

LIU, Y., SWAMINATHAN, A., AGARWAL, A. and BRUNSKILL, E. (2020). Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202* .

LOVEJOY, W. S. (1991). A survey of algorithmic methods for partially observed markov decision processes. *Annals of Operations Research* **28** 47–65.

MASTOURI, A., ZHU, Y., GULTCHIN, L., KORBA, A., SILVA, R., KUSNER, M., GRETTON, A. and MUANDET, K. (2021). Proximal causal learning with kernels: Two-stage estimation and moment restriction. In *International Conference on Machine Learning*. PMLR.

MIAO, W., GENG, Z. and TCHETGEN TCHETGEN, E. J. (2018a). Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* **105** 987–993.

MIAO, W., SHI, X. and TCHETGEN, E. T. (2018b). A confounding bridge approach for double negative control inference on causal effects. *arXiv preprint arXiv:1808.04945* .

MUNDHENK, M., GOLDSMITH, J., LUSENA, C. and ALLENDER, E. (2000). Complexity of finite-horizon Markov decision process problems. *Journal of the ACM (JACM)* **47** 681–720.

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

OBERST, M. and SONTAG, D. (2019). Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*. PMLR.

PAPADIMITRIOU, C. H. and TSITSIKLIS, J. N. (1987). The complexity of markov decision processes. *Mathematics of operations research* **12** 441–450.

PEARL, J. (2009). *Causality*. Cambridge university press.

RASHIDINEJAD, P., ZHU, B., MA, C., JIAO, J. and RUSSELL, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *arXiv preprint arXiv:2103.12021* .

ROCKAFELLAR, R. T. and WETS, R. J.-B. (2009). *Variational analysis*, vol. 317. Springer Science & Business Media.

SHALEV-SHWARTZ, S., SHAMMAH, S. and SHASHUA, A. (2016). Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295* .

SHAPIRO, A., DENTCHEVA, D. and RUSZCZYNSKI, A. (2021). *Lectures on stochastic programming: modeling and theory*. SIAM.

SHI, C., UEHARA, M. and JIANG, N. (2021). A minimax learning approach to off-policy evaluation in partially observable markov decision processes. *arXiv preprint arXiv:2111.06784* .

SHI, C., WANG, X., LUO, S., ZHU, H., YE, J. and SONG, R. (2022). Dynamic causal effects evaluation in a/b testing with a reinforcement learning framework. *Journal of the American Statistical Association* 1–29.

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

SUN, P., KRETZSCHMAR, H., DOTIWALLA, X., CHOUARD, A., PATNAIK, V., TSUI, P., GUO, J., ZHOU, Y., CHAI, Y., CAINE, B. ET AL. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

SUTTON, R. S. and BARTO, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

TCHETGEN, E. J. T., YING, A., CUI, Y., SHI, X. and MIAO, W. (2020). An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982* .

TENNENHOLTZ, G., SHALIT, U. and MANNOR, S. (2020). Off-policy evaluation in partially observable environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34.

UEHARA, M., IMAIZUMI, M., JIANG, N., KALLUS, N., SUN, W. and XIE, T. (2021). Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981* .

UEHARA, M. and SUN, W. (2021). Pessimistic model-based offline rl: Pac bounds and posterior sampling under partial coverage. *arXiv e-prints* arXiv–2107.

VLASSIS, N., LITTMAN, M. L. and BARBER, D. (2012). On the computational complexity of stochastic controller optimization in POMDPs. *ACM Transactions on Computation Theory (TOCT)* **4** 1–8.

WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge University Press.

XIE, T., CHENG, C.-A., JIANG, N., MINEIRO, P. and AGARWAL, A. (2021). Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems* **34**.

XIE, T. and JIANG, N. (2020). Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*. PMLR.

XIONG, Y., CHEN, N., GAO, X. and ZHOU, X. (2021). Sublinear regret for learning pomdps. *arXiv preprint arXiv:2107.03635* .

YAN, Y., LI, G., CHEN, Y. and FAN, J. (2022). The efficacy of pessimism in asynchronous q-learning. *arXiv preprint arXiv:2203.07368* .

YIN, M., DUAN, Y., WANG, M. and WANG, Y.-X. (2022). Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804* .

YIN, M. and WANG, Y.-X. (2021). Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems* **34**.

YU, T., KUMAR, A., RAFAILOV, R., RAJESWARAN, A., LEVINE, S. and FINN, C. (2021). Combo: Conservative offline model-based policy optimization. *Advances in Neural Information Processing Systems* **34**.

ZANETTE, A. (2021). Exponential lower bounds for batch reinforcement learning: Batch rl can be exponentially harder than online rl. In *International Conference on Machine Learning*. PMLR.

ZANETTE, A., WAINWRIGHT, M. J. and BRUNSKILL, E. (2021). Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems* **34**.

ZHAN, W., HUANG, B., HUANG, A., JIANG, N. and LEE, J. D. (2022). Offline reinforcement learning with realizability and single-policy concentrability. *arXiv preprint arXiv:2202.04634* .

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