Off-policy Evaluation in Infinite-horizon Reinforcement Learning with Latent Confounders

Andrew Bennett
Cornell University
awb222@cornell.edu

Nathan Kallus
Cornell University
kallus@cornell.edu

Lihong Li
Amazon
llh@amazon.com

Ali Mousavi
Work Done at Google
ali.mousavi1988@gmail.com

Abstract

Off-policy evaluation (OPE) in reinforcement learning is an important problem in settings where experimentation is limited, such as education and healthcare. But, in these very same settings, observed actions are often confounded by unobserved variables making OPE even more difficult. We study an OPE problem in an infinite-horizon, ergodic Markov decision process with unobserved confounders, where states and actions can act as proxies for the unobserved confounders. We show how, given only a latent variable model for states and actions, policy value can be identified from off-policy data. Our method involves two stages. In the first, we show how to use proxies to estimate stationary distribution ratios, extending recent work on breaking the curse of horizon to the confounded setting. In the second, we show optimal balancing can be combined with such learned ratios to obtain policy value while avoiding direct modeling of reward functions. We establish theoretical guarantees of consistency, and benchmark our method empirically.

1 Introduction

A fundamental question in offline reinforcement learning (RL) is how to estimate the value of some target evaluation policy, defined as the long-run average reward obtained by following the policy, using data logged by running a *different* behavior policy. This question, known as off-policy evaluation (OPE), often arises in applications such as healthcare, education, or robotics, where experimenting with running the target policy can be expensive or even impossible, but we have data logged following business as usual or current standards of care. A central concern using such passively observed data is that observed actions, rewards, and transitions may be *confounded* by unobserved variables, which can bias standard OPE methods that assume no unobserved confounders, or equivalently that a standard Markov decision process (MDP) model holds with fully observed state.

Consider for example evaluating a new smart-phone app to help people living with type-1 diabetes time their insulin injections by monitoring their blood glucose level using some wearable device. Rather than risking giving bad advice that may harm individuals, we may consider first evaluating our injection-timing policy using existing longitudinal observations of individuals' blood glucose levels over time and the timing of insulin injections. The value of interest may be the long-run average deviation from ideal glucose levels. However, there may in fact be events not recorded in the data, such as food intake and exercise, which may affect both the timing of injections and blood glucose. Unfortunately, most previously proposed methods for OPE in RL setting do not account for such confounding, so if they are used for analysis the results may be biased and misleading.

In this work, we study OPE in an infinite-horizon, ergodic MDP with unobserved confounders, where states and actions can act as proxies for the unobserved confounders. We show how, given only a latent variable model for states and actions, the policy value can be identified from off-policy data. We provide an optimal balancing (Bennett & Kallus, 2019) algorithm for estimating the policy value while avoiding direct modeling of reward functions, given an estimate of the stationary distribution ratio of states and an identified model of confounding. In addition, we provide an algorithm for estimating the stationary distribution ratio of states in the presence of unobserved confounders, by extending recent work on infinite-horizon OPE (Liu et al., 2018) and efficiently solving conditional moment matching problems (Bennett et al., 2019). On the theory side, we establish statistical consistency under the assumption of iid confounders, and provide error bounds for our method in close-to-iid settings. Finally, we demonstrate that our method achieves strong empirical performance compared with several causal and non-causal baselines.

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

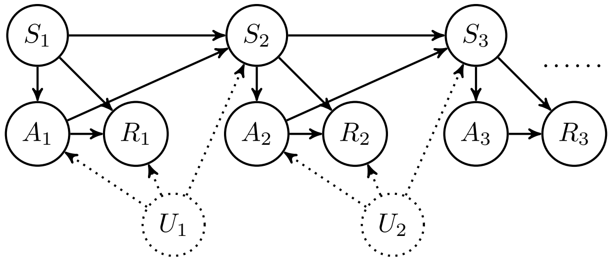

Figure 2: Graphical representation of the MDPUC model, in which action selection, state transition, and reward emission are confounded *at every step*.

Notation We use uppercase letters such as S and X to denote random variables, and lowercase ones to denote nonrandom quantities. The set of positive integers is $\mathbb{N}$, and for any $n \in \mathbb{N}$ we use $[n]$ to refer to the set $\{1, \dots, n\}$. We denote by $\| \cdot \|_p$ the usual functional norm, defined as $\|f\|_p = \mathbb{E}[|f(X)|^p]^{1/p}$, where the measure is implicit from context. Furthermore we denote as $\mathcal{L}_p$ the space of functions with finite $\| \cdot \|_p$-norm. We denote by $N(\epsilon, \mathcal{F}, \| \cdot \|)$ the ϵ-covering number of $\mathcal{F}$ under metric $\| \cdot \|$, and the corresponding bracketing number by $N_{[]}(\epsilon, \mathcal{F}, \| \cdot \|)$. Finally, for any random variable sequence $\{Q_1, Q_2, \dots\}$, we use the notation $Q_{l:u}$ as shorthand for $(Q_l, Q_{l+1}, \dots, Q_u)$.

2 Problem Setting

We consider the Markov Decision Process with Unmeasured Confounding, or MDPUC (Zhang & Bareinboim, 2016), which is a confounded generalization of a standard Markov decision process (MDP). An MDPUC is specified by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{U}, P_T, \mathcal{R}, P_0)$, where $\mathcal{S}$ is the finite state space, $\mathcal{A} = [m]$ the action space, $\mathcal{U}$ the confounder space, $P_T(s' \mid s, a, u)$ the probability of transitioning to state s' from state s given action a and confounder u, $\mathcal{R}(s, a, u)$ the reward distribution given action a was taken in state s with confounder u, and P_0 the distribution over starting states. We also define $\mu_a(s, u) = \mathbb{E}[R(s, a, u)]$, where $R(s, a, u)$ is any random variable distributed according to $\mathcal{R}(s, a, u)$, and we use S' to refer to the state succeeding state S in a trajectory, Z to refer to the triplet (S, A, S'), and X to refer to the pair (Z, U). An important assumption we make here is that the confounder values U at each time step are iid, which differentiates the MDPUC setting from the more general POMDP setting. An example of this setting may be our diabetes problem from section 1, where S corresponds to blood glucose levels, A corresponds to insulin injection decisions, R is based on maintaining safe blood glucose levels, and U corresponds to the exogenous unmeasured events such as food intake or exercise.[1]

We assume access to $N \geq 1$ trajectories of off-policy data, of lengths $T_1, \dots, T_N$. At each time step of a trajectory we assume that we observe the state S, the action that was taken in that state A, and the corresponding reward that was received R. Importantly, we do *not* observe the corresponding confounder value U. We assume that each trajectory was logged from a common behavior policy π_b, which depends on the confounders, where $\pi_b(a \mid s, u)$ gives the probability that π_b takes action a given state s and confounder u. Note that although we assume our data is collected from separate trajectories, for brevity we will index our data by concatenating these trajectories together and using indices $i \in [n]$, where $n = \sum_{i=1}^{N} T_i$, and we denote the observed data by $\mathfrak{D} = \{Z_i, R_i\}_{i \in [n]}$.

Our task is to estimate the value of some fixed evaluation policy π_e, which follows the same semantics as π_b, and whose actions may optionally depend on the confounders U (for simplicity, even in the case that its actions depend on S only, we still use the notation $\pi_e(a \mid s, u)$) We make the following ergodicity and mixing assumptions about the behavior and evaluation policies.

Assumption 1 (Ergodicity). *For some $2 < p \leq \infty$ we have The Markov chain of X values under each of π_b and π_e is ergodic, and , the chain of X values under π_b is stationary. Furthermore, $\sum_{k=1}^{\infty} k^{2/(p-2)} \beta(k) < \infty$, where $\beta(k)$ are the β-mixing coefficients of the Markov chain of X values induced by π_b.*

Assumption 1 uses β-mixing coefficients, which quantify how close to independent X values k steps removed are in the Markov chain, with coefficients of zero implying independence. In our stationary Markovian setting, these are defined according to the expectated total variation distance between the marginal distribution of X_{k+1} and its conditional distribution given X_1; that is

$$\beta(k) = \mathbb{E}[\sup_{B \in \sigma(X_{k+1})} |P(B \mid X_1) - P(B)|] \,,$$

where $\sigma(X)$ denotes the σ-algebra generated by X. Overall, the assumption implies that the X values obtained from each policy have a unique stationary distribution, where the dependence between far removed elements is sufficiently weak, and under π_b all values follow this stationary distribution. Note that this is a very standard kind of assumption, with similar assumptions in most prior work on OPE in infinite-horizon settings that addresses the curse of horizon (Liu et al., 2018; Kallus & Uehara, 2019b). Without a similar assumption the kind of analysis we perform would be impossible, because either stationary distributions would not exist, or we would not be able to control the interactions between distant data points in order to bound error terms.

We let $\mathbb{E}_b$ and $\mathbb{E}_e$ denote expectations taken with respect to these stationary distributions, and assume that probability statements refer to the stationary distribution under π_b where

[1] Although the confounders are likely not perfectly iid, this modelling approximation may be justified for instance if we can approximately model the confounding events by a Poisson process.

not specified. In addition, we will use the notation $d(Q)$ to denote the stationary density ratio under π_e versus π_b, for any random variable Q that is measurable with respect to X.[2] Given this, we define the *value* of π_e to be

$$v(\pi_e) = \mathbb{E}_e[\mu_A(S,U)] \,.$$

Finally, we make the following basic regularity assumptions.

Assumption 2 (State Visitation Overlap)**.** $\|d(S)\|_\infty < \infty$.

Assumption 3 (Bounded Reward Moments)**.** *For each a, $\mu_a(S,U)$ is uniformly bounded almost surely. In addition, $\mathbb{V}[R \mid Z] \leq \sigma^2$ and $\mathbb{V}[R \mid Z, U] \leq \sigma^2$ almost surely, for some constant σ^2.*

Assumption 2 is a very fundamental assumption, and is analogous to treatment overlap assumptions that are ubiquitous in treatment effect estimation (e.g., Murphy et al., 2001). Fundamentally, in order to be able to perform causal inference without strong structural assumptions, we require overlap between the observed distribution of data, and the distribution that would be obtained under intervention. Assumption 3 is a very simple assumption that is trivially satisfied when rewards are bounded, as in most realistic applications. We chose, however, to include this more complex assumption rather than assuming bounded rewards, in order to make our theory more general.

Finally, we note that the problem of OPE that we are considering is distinct from the related problem of *policy learning*, where given the same kind of logged data the task is to choose a policy π within some class of policies Π that (approximately) maximizes $v(\pi)$. We do not explicitly consider the problem of policy learning under unmeasured confounding in this paper, however we note that our proposed methods for OPE could be used to construct an objective function for policy learning. We leave additional study of this related problem, and analysis of the impact of unmeasured confounding on it, to future work.

3 Related Work

The infinite-horizon OPE problem has received fast-growing interest recently (Liu et al., 2018; Gelada & Bellemare, 2019; Kallus & Uehara, 2019b; Nachum et al., 2019a; Mousavi et al., 2020; Uehara et al., 2020; Liu et al., 2020; Dai et al., 2020). Most of these approaches are based on some form of moment matching condition, derived from the stationary distribution of the corresponding Markov chains, and can avoid the exponential growth of variance in typical importance sampling methods (Liu et al., 2018). Our work extends this research to a more general setting with unobserved confounders. Similar to our work, Tennenholtz et al. (2020) has addressed OPE under unmeasured confounding in the POMDP setting, however their work relies on complex invertibility assumptions and is limited to tabular settings. In addition there is a tangential line of work investigating the limitations of OPE under unmeasured confounding in nonparametric settings and constructing partial identification bounds (Kallus & Zhou, 2020; Namkoong et al., 2020), which differs from our focus on specific settings where the model of confounding is identifiable and therefore so is the policy value. Furthermore, OPE under unmeasured confounding has been studied in contextual bandit settings (Bennett & Kallus, 2019), which may be viewed as a special case of our problem where states are generated iid in every step.

Related to the *evaluation* problem is policy *learning*, where the goal is to interact with an unknown environment to optimize the policy. The partially observable MDP (POMDP) is a classic model for sequential decision making with unobserved state (Kaelbling et al., 1998), and has been extensively studied (Spaan, 2012; Azizzadenesheli et al., 2016). More recently, a few authors have applied counterfactual reasoning techniques to multi-armed bandits and, more generally, RL (Bareinboim et al., 2015; Zhang & Bareinboim, 2016; Lu et al., 2018; Buesing et al., 2019). While evaluation might appear simpler than learning, OPE methods only have access to a fixed set of data and cannot explore. This restriction leads to different challenges in algorithmic development that are tackled by our proposed method.

Finally, another related area of research is on using proxies for true confounders (Wickens, 1972; Frost, 1979). Much of this work involves fitting and using latent variable models for confounders, or studying sufficient conditions for identification of these latent variable models (Cai & Kuroki, 2008; Wooldridge, 2009; Pearl, 2012; Kuroki & Pearl, 2014; Edwards et al., 2015; Louizos et al., 2017; Kallus et al., 2018). This body of research is complementary to our work, since we propose an estimator that uses a latent variable model for confounders, but do not study how to fit it.

4 Theory for Optimally Weighted Policy Evaluation

In this work, we consider generic weighted estimators of the form

$$\hat{\tau}_W = \frac{1}{n}\sum_{i=1}^{n} W_i R_i \,, \tag{1}$$

where $W = W_{1:n}$ is any vector of weights that is measurable with respect to $Z_{1:n}$. Inspired by Kallus (2018) and Bennett & Kallus (2019), we proceed by deriving an upper-bound for the risk of policy evaluation. First, we observe

[2] That is, for any such Q, we define $d(Q)$ such that $\mathbb{E}_e[g(Q)] = \mathbb{E}_b[d(Q)g(Q)]$ for any measurable function g. Note that this involves slight abuse of notation since the function d depends on the random variable Q.

that the value of π_e is given by

$$\begin{aligned} v(\pi_e) &= \sum_{a=1}^{m} \mathbb{E}_e[\pi_e(a \mid S, U)\mu_a(S,U)] \\ &= \sum_{a=1}^{m} \mathbb{E}_b[d(S)\pi_e(a \mid S, U)\mu_a(S,U)] \,, \end{aligned}$$

where the second equality follows from the observation that $d(S,U) = d(S)$ under the iid confounder assumption. In addition, it is easy to verify that $\mathbb{E}_b[WR] = \sum_{a=1}^{m} \mathbb{E}_b[W\delta_{Aa}\mu_a(S,U)]$. It suggests that if we knew $U_{1:n}$, the bias of balanced policy evaluation could be approximated by

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{m}(W_i\delta_{A_i a} - d(S_i)\pi_e(a \mid S_i, U_i))\mu_a(S_i, U_i),$$

which motivates the following theorem.

Theorem 1. *For any vector W, vector-valued function $g = (g_1, \ldots, g_m)$, and constant λ, define*

$$\begin{aligned} f_{ia} &= W_i\delta_{A_i a} - d(S_i)\pi_e(a \mid S_i, U_i) \,, \\ B(W,g) &= \frac{1}{n}\sum_{i=1}^{n}\sum_{a=1}^{m}\mathbb{E}[f_{ia}g_a(S_i,U_i) \mid Z_i] \,, \\ J_\lambda(W,g) &= B(W,g)^2 + \frac{\lambda}{n^2}\|W\|^2 \,. \end{aligned}$$

Then, if $\lambda \geq 4\sigma^2$ and $J_\lambda(W,\mu) = O_p(r_n)$, where $\mu = (\mu_1, \ldots, \mu_m)$ are the true mean reward functions, it follows from assumptions 1 to 3 that

$$\hat{\tau}_W = v(\pi_e) + O_p(\max(n^{-1/2}, r_n^{1/2})) \,.$$

This result suggests finding weights W in eq. (1) that minimize $\sup_{g\in\mathcal{G}} J_\lambda(W,g)$ for some vector-valued function class $\mathcal{G}$, since if $\mu \in \mathcal{G}$ and we can minimize this upper bound uniformly over $\mathcal{G}$ at an $O_p(1/n)$ rate, then $\hat{\tau}_W$ is $O_p(1/\sqrt{n})$-consistent for $v(\pi_e)$.

Next, we describe a category of function classes for which the above $O_p(1/\sqrt{n})$ convergence is achievable.

Definition 1 (Balancing-regular Class)**.** *Let $\mathcal{G}$ be some normed vector valued function class, and for $a \in [m]$ define*

$$\begin{aligned} &\mathcal{G}^* = \{g/\|g\| : g \in \mathcal{G}\} \\ &\mathcal{G}_a^* = \{g_a : \exists(g_1', \ldots, g_m') \in \mathcal{G}^* \text{ with } g_a = g_a'\} \,. \end{aligned}$$

*Then we say that $\mathcal{G}$ is p-*balancing-regular *if it satisfies the following properties:*

1. *$\mathcal{G}$ and $\mathcal{G}^*$ are compact.*
2. *$\mathcal{G}$ is convex.*
3. *$g \in \mathcal{G} \iff -g \in \mathcal{G}$*
4. *$g_a(s,u)$ is continuous in g for every $s \in \mathcal{S}$ and $u \in \mathcal{U}$, and is continuous in s and u for every $g \in \mathcal{G}$.*
5. *There exists some constant G such that $\|g_a\|_\infty \leq G$ for every $g_a \in \mathcal{G}_a^*$ and $a \in [m]$.*
6. *$\int_0^\infty \sqrt{\log N_{[]}(\epsilon, \mathcal{G}_a^*, \mathcal{L}_p)}d\epsilon < \infty$ for each $a \in [m]$ and every possible joint measure on S and U,*

where $N_{[]}(\epsilon, \mathcal{F}, d)$ denotes the ϵ-bracketing number of function set $\mathcal{F}$ under metric d.

We will consider functions classes that are p-balancing-regular, where p is same constant as from assumption 1. It is easy to show that many commonly considered function classes are p-balancing-regular. In particular, we provide the following lemma, which justifies that this holds for a variety of Reproducing Kernel Hilbert Spaces (RKHSs).

Lemma 1. *Let K be a symmetirc, PSD, C^∞-smooth and and bounded kernel, and let $\|g\|^2 = \sum_{a=1}^{m}\|g_a\|_K^2$, where $\|\cdot\|_K$ is the RKHS norm with kernel K. Then for any $\gamma > 0$, the function class $\mathcal{G}_{K,\lambda} = \{g : \|g\| \leq \gamma\}$ is p-balancing-regular for every $p > 0$.*

Finally, we make the following assumption to avoid the pathological situation where $\mathbb{E}[g_A(S,U) \mid Z] = 0$ almost surely for some $g \neq 0$, in which case $B(W,g) = B(W',g)$ for any $W, W' \in \mathbb{R}^n$ and bias cannot be controlled.

Assumption 4 (Non-degeneracy)**.**

$$\sup_{g\in\mathcal{G}^*} P(\mathbb{E}[g_A(S,U) \mid Z] = 0) < 1.$$

Note that this is a joint assumption on the class $\mathcal{G}$ and the data generating process, and is similar to identifiability conditions in other causal inference works with latent variable such as in Miao et al. (2018); it can be seen as the assumption that any $\mu, \mu' \in \mathcal{G}$ with $\mu \neq \mu'$ would induce a different observed distribution of data.

Given this additional assumption and our p-balancing-regular definition, we can now present our next core theorem, which justifies that we can uniformly minimize the adversarial objective J_λ.

Theorem 2. *Given assumptions 1 and 4, and assuming that $\mathcal{G}$ is p-balancing-regular, where p is the same constant as in assumption 1, we have*

$$\inf_{W\in\mathbb{R}^n}\sup_{g\in\mathcal{G}} J_\lambda(W,g) = O_p(1/n) \,.$$

4.1 Oracle Consistency

We now present the most basic consistency result, which considers the oracle setting where we have an oracle for the conditional distribution of U given Z, and for the state density ratio d. This implies that for any W

and g, we can compute $J_\lambda(W, g)$ exactly. Let $W^* = \arg\min_W \sup_{g \in \mathcal{G}} J_\lambda(W, g)$. Then we have the following oracle consistency theorem.

Theorem 3. *Given assumptions 1 to 4, then as long as $\lambda \geq 4\sigma^2$ and $\mu \in \mathcal{G}$, we have $\hat{\tau}_{W^*} = v(\pi_e) + O_p(n^{-1/2})$.*

The theorem follows by chaining together theorems 1 and 2, and noting that $\mu \in \mathcal{G}$ implies that $J_\lambda(W, \mu) \leq \sup_{g \in \mathcal{G}} J_\lambda(W, \mu)$. We note that, although a condition of this result is that $\lambda \geq 4\sigma^2$, this is without loss of generality for norm-bounded function classes, since replacing λ with $4\sigma^2$ and $\{g : \|g\| \leq \gamma\}$ with $\{g : \|g\| \leq 2\sigma\gamma/\sqrt{\lambda}\}$ will give an equivalent optimization problem in W.[3] However, the condition that $\mu \in \mathcal{G}$ is fundamental.

4.2 Sensitivity to Nuisance Estimation Error and Model Misspecification

Next we extend our theory to more realistic settings, and consider the effects of estimation errors and non-iid confounding. We present simplified results here for the common case where π_e is measurable with respect to S only, with corresponding results for the general case where π_e can also depend on U presented in appendix C. For this analysis, we let some normed function class $\mathcal{F}$ be given. Then, for any measures p and q on $\mathcal{U}$ we define the integral probability metric

$$D_\mathcal{F}(p, q) = \sup_{\|f\|_\mathcal{F} \leq 1} |\int f(u) dp(u) - \int f(u) dq(u)| \,.$$

Examples of this include total variation distance, where $\|f\|_\mathcal{F} = \|f\|_\infty$, the maximum mean discrepancy where $\|f\|_\mathcal{F}$ is given by some RKHS norm, and Wasserstein distance, where $\|f\|_\mathcal{F}$ is given by the Lipschitz norm.

We first address the issue that the adversarial objective considered above depends on the conditional density of U given Z, and the state density ratio d. In practice these both would usually need to be estimated from data. Let $\varphi(z)$ and $\hat{\varphi}(z)$ denote the true and estimated conditional distribution of U respectively given $Z = z$, let $\hat{d}$ be the estimated state density ratio. In addition let $\hat{J}_\lambda(W, g)$ denote the objective using $\hat{\varphi}$ and $\hat{d}$ in place of φ and d, and let $W^* = \arg\min_W \sup_{g \in \mathcal{G}} \hat{J}(W, g)$.

Lemma 2. *Suppose that there exists some constant $F < \infty$ such that for every $g \in \mathcal{G}$ and $a \in [m]$ we have $\|g_a\|_\mathcal{F} \leq F$ and $\|\mu_a\|_\mathcal{F} \leq F$. Suppose in addition that $D_\mathcal{F}(\varphi(Z_i), \hat{\varphi}(Z_i)) = O_p(r_n)$, and $|d(S_i) - \hat{d}(S_i)| = O_p(r_n)$, for every $i \in [n]$. Then, given assumptions 1 to 4 and assuming $\mu \in \mathcal{G}$ and $\mathcal{G}$ is p-balancing-regular, for the p defined in assumption 1, we have*

$$\hat{\tau}_{W^*} = v(\pi_e) + O_p(\max(n^{-1/2}, n^{1/2} r_n^2)) \,.$$

Note that r_n is not some fixed rate; lemma 2 simply says that if its conditions hold with *any* given rate r_n, then we obtain the corresponding bound on the rate of convergence for $\hat{\tau}_{W^*}$. In particular, this implies that our methodology will be consistent as long as φ and d are estimated at a $o_p(n^{-1/4})$ rate, and also that we can still obtain $O_p(n^{-1/2})$-consistency if φ and d are estimated at a $O_p(n^{-1/2})$ rate. If we are willing to assume a correctly-specified parametric model for φ and fit it via maximum likelihood estimation, then the $O_p(n^{-1/2})$ rate easily follows under mild differentiability assumptions.[4] We discuss this in further detail in appendix D.1, as well as other approaches and corresponding rates for estimating φ. In addition, we discuss the estimation of d in section 5.1. We also note that by assumption 3 and the definition of p-balancing-regular, the condition that a finite constant F exists is immediately follows in the case that we use the ∞-norm for $\mathcal{F}$. However, the presentation of the above theorem for a general $\mathcal{F}$ makes our theory more general.

Next, we consider minor violations in the iid confounder assumption of the MDPUC model. Specifically, we consider an alternate model where U values form a Markov chain. Under this alternate model, we provide the following theorem bounding the squared error.

Theorem 4. *Suppose that the conditions of lemma 2 hold, and $\|d(S, U)\|_\infty < \infty$. In addition let φ_i and φ_i^* denote the conditional densities of U_i given Z_i and $Z_{1:n}$, let $b = \max_a \|\mu_a\|_\infty$, and let $c = \sqrt{2}F(1 + \|W^*\|^2/n)^{1/2}$. Then we have $(\hat{\tau}_{W^*} - v(\pi_e))^2 = \epsilon^2 + O_p(\max(1/n, n r_n^4))$, where*

$$|\epsilon| \leq c \left(\frac{1}{n} \sum_{i=1}^n D_\mathcal{F}(\varphi_i, \varphi_i^*)^2 \right)^{1/2} + b \|d(S, U) - d(S)\|_2 \,.$$

We note that in the iid confounder case $\varphi_i = \varphi_i^*$ and $d(S, U) = d(S)$, so the first two terms disappear, and the result reduces to that of lemma 2. In addition under assumption 3, the constant b must be finite. Therefore, theorem 4 allows us to bind the asymptotic bias[5] in "near-iid" settings, where the terms $D_\mathcal{F}(\varphi_i, \varphi_i^*)$ and $\|d(S) - d(S, U)\|_2$ are small. Note that the actual magnitude of this asymptotic bias will be problem specific and depend on the problem parameters. Rather, theorem 4 establishes that the smaller the violation of our iid assumption is, the smaller the resulting asymptotic bias will be.

4.3 Consistency under $\mathcal{G}$ Misspecification

All of our theory so far has been conditioned on the strong and untestable assumption that our function class $\mathcal{G}$ is correctly specified; that is, $\mu \in \mathcal{G}$. We now generalize our theory to the case where we use a *universally approximating*

[3] This is because $\sup_{\|g\| \leq \gamma} J_\lambda(W, g)$ is identical to $(\lambda/4\sigma^2) \sup_{\|g\| \leq 2\sigma\gamma/\sqrt{\lambda}} J_{4\sigma^2}(W, g)$.

[4] We emphasize, however, that our theory *does not* necessarily assume a correctly-specified parametric model for φ. This is simply presented as an example of how we might justify a given rate of convergence of $\hat{\varphi}$.

[5] That is, the limiting bias as $n \to \infty$.

function class. Specifically, we will consider a sequence of function classes $\{\mathcal{G}_1, \mathcal{G}_2, \ldots\}$ to be universally approximating if, for any continuous function g, we have

$$\lim_{i \to \infty} \sup_{g' \in \mathcal{G}_i} \|g' - g\|_\infty = 0\,.$$

This property holds for many commonly used function classes, such as Gaussian RKHSs with shrinking scale parameter on the kernel (Mendelson, 2003). Given this definition, we present the following lemma.

Lemma 3. *Suppose that the sequence $\{\mathcal{G}_i\}$ is universally approximating, such that $\mathcal{G}_i$ is p-balancing-regular for each i. Then given all the assumptions of theorem 4 except that $\mu \in \mathcal{G}$, and instead just assuming that μ is continuous, there exists some non-decreasing sequence of integers i_n such that the weighted estimator using $W^* = \arg\min_W \sup_{\mathcal{G}_{i_n}} \hat{J}_\lambda(W, g)$ satisfies theorem 4 with the O_p term replaced with $o_p(1)$.*

We note that the sequence i_n for which this lemma is satisfied depends on the specific sequence of function classes $\{\mathcal{G}_i\}$; this lemma just ensures that there is *some* rate at which we can grow the complexity of $\mathcal{G}$ and still obtain consistency. As detailed in our proof in the appendix, the optimal rate i_n depends both on the rate at which $\sup_{g \in \mathcal{G}_i} \|g - \mu\|_\infty$ converges to zero, and the rate at which the constant factor in the O_p term from theorem 4 grows with $\mathcal{G}_i$. We leave the question of calculating the optimal i_n for particular universally approximating sequences $\{\mathcal{G}_i\}$ to future work.

4.4 Discussion of Assumptions

We acknowledge that much of our consistency theory depends on strong and untestable assumptions. In particular, we assume either explicitly or implicitly throughout most of our theory that confounder values are iid, and that we can estimate the posterior confounder distributions $\varphi(Z)$ at reasonably fast rates. We believe that these assumptions are reasonable to make for various reasons.

First, we argue the iid assumption is necessary for statistic tractability, since otherwise the joint posterior distribution of the confounder values $U_{1:n}$ could depend arbitrarily on $Z_{1:n}$,[6] which in general would not factorize nicely, and in any corresponding estimator to ours we would have to integrate with respect to this joint distribution. This would likely make the resulting algorithm both computationally and statistically intractable. Furthermore, we believe that in many applications this assumption should hold at least approximately, such as in our previous diabetes management example, and more generally wherever the correlation between temporally-distant confounder values is weak. In such settings, we can bound the resulting bias by theorem 4. In addition, in general it is impossible to perform consistent policy evaluation under unmeasured confounding without some fairly strong assumptions, as established by the non-identifiability of the policy value in general in this setting (e.g., Kallus & Zhou, 2020). Hence, while these assumptions may seem unsatisfying, we argue that strong assumptions such as these are necessary. Furthermore, these assumptions are weaker than those of the standard OPE setting, where there is no unmeasured confounding, since our work subsumes this as a special case.

5 Methodology

We now discuss how to implement the optimal balancing estimator analyzed in section 4. It can be done in three steps: (1) estimating the conditional distribution of U given Z (denoted by φ); (2) estimating d; and (3) calculating $W^* = \arg\min_W \sup_{g \in \mathcal{G}} \hat{J}_\lambda(W, g)$. The first has been studied extensively, so we focus only on the last two steps.

5.1 Estimating the Stationary Density Ratio

Here, we pose learning the stationary density ratio $d(S)$ as a conditional moment matching problem. Similarly to Liu et al. (2018), we can identify d via a set of moment restrictions, as follows.

Theorem 5. *Let $\beta(z) = \mathbb{E}[\pi_e(A \mid S, U)/\pi_b(A \mid S, U) \mid Z = z]$. Then under assumption 1, the stationary density ratio $d(S)$ is the unique function satisfying the regular moment condition $\mathbb{E}[d(S)] = 1$, as well as the conditional moment restriction $d(S') = \mathbb{E}[d(S)\beta(Z) \mid S']$ for almost-everywhere S'.*

We note that the conditional moment restriction above is equivalent to $\mathbb{E}[h(S')d(S)\beta(Z)] = 0$ for every function h. There is a variety of work on solving such conditional moment restrictions; see *e.g.* Carrasco et al. (2007); Muandet et al. (2019); Bennett et al. (2019); Dikkala et al. (2020); Bennett & Kallus (2020) and citations therein. In particular, following the efficient variational approaches of Bennett et al. (2019); Bennett & Kallus (2020), we propose to estimate d by solving a smooth-game optimization problem. Let function classes $\mathcal{H}$ and $\mathcal{D}$ be given, and let $\tilde{d}$ by some prior estimate of d, which might come from a previous generalized method of moments (GMM) estimate or some other methodology. In addition, define

$$m(Z; d, h, c) = h(S')(d(S)\beta(Z) - d(S')) + c(d(S) - 1)$$

$$U_n(d, h, c) = \frac{1}{n}\sum_{i=1}^n m(Z_i; d, h, c) - \frac{1}{4n}\sum_{i=1}^n m(Z_i; \tilde{d}, h, c)^2.$$

Our proposed estimator takes the form

$$\hat{d} = \arg\min_{d \in \mathcal{D}} \sup_{h \in \mathcal{H}, |c| \le \lambda_c} U_n(d, \tilde{d}, h, c)\,. \qquad (2)$$

[6] More accurately, the set of all confounder values within each distinct trajectory could depend arbitrarily on the entire set of observed data within that same trajectory.

The motivation of this form of estimator is that, if we choose $\mathcal{H} = \text{span}(\{h_1, \ldots, h_k\})$, and $\lambda_c = \infty$, then eq. (2) is equivalent to a standard efficiently weighted GMM estimator (Hansen, 1982) on the standard moment conditions $\mathbb{E}[d(S)] = 1$, and $\mathbb{E}[h_i(S')d(S)\beta(Z)] = 0$ for $i \in [k]$. This follows from a generalization of (Bennett et al., 2019, Lemma 1), which we explain in more detail in appendix E. Therefore, we can view eq. (2) as a regularized analogue of efficiently weighted GMM on an infinite set of moment conditions. Given this, in the case that $\mathcal{D}$ is parametric and we choose $\mathcal{H}$ and λ_c as above, then $O_p(n^{-1/2})$-convergence of $\hat{d}$ can be established under standard regularity conditions for GMM estimators (Hansen, 1982). We further discuss this, as well as known results for general $\mathcal{D}$, in appendix D.2.

In practice, we can start with an initial guess for $\tilde{d}$ (such as $\tilde{d}(s) = 1 \; \forall s$), and then iteratively solve eq. (2) with $\tilde{d}$ being the previous solution. Furthermore, for our experiments we choose to use norm-bounded RKHSs for $\mathcal{H}$ and $\mathcal{D}$,[7] which allows the optimization to be performed analytically, as dictated by the following theorem.

Theorem 6. *Suppose we observe n_s distinct states in our dataset, which we denote $\{s_1, \ldots, s_{n_s}\}$. Let Lagrangian regularization hyperparameters $\lambda_h, \lambda_k > 0$ be given, let K_H and K_D denote the reproducing kernels for $\mathcal{H}$ and $\mathcal{D}$ respectively, and let k_H and k_D the corresponding kernel matrices on the states $s_1, \ldots, s_{n_s}$. Furthermore, define*

$$\begin{aligned}
q_i &= [K_H(s_1, S_i')\tilde{d}(S_i)\beta(Z_i), \\
&\qquad \ldots, K_H(s_{n_s}, S_i')\tilde{d}(S_i)\beta(Z_i), \tilde{d}(S_i) - 1]^T \\
Q &= \frac{1}{n}\sum_{l=1}^{n} q_l q_l^T + \text{BlockDiag}(\lambda_h k_H, \lambda_c) \\
\Omega_{i,j} &= \frac{1}{n}\sum_{l=1}^{n} K_H(s_i, S_l')K_D(s_j, S_l)\beta(Z_l) \\
&\qquad \forall i, j \in [n_s] \\
\Omega_{n_s+1,j} &= \frac{1}{n}\sum_{l=1}^{n} K_D(s_j, S_l) \quad \forall j \in [n_s] \\
\omega &= [0, \ldots, 0, 1]^T \\
b &= (\Omega^T Q^{-1}\Omega + \lambda_d k_D)^{-1}\Omega^T Q^{-1}\omega \,.
\end{aligned}$$

Then, the solution to eq. (2) *using RKHS balls for $\mathcal{H}$ and $\mathcal{D}$ is given by*

$$\hat{d}(s) = \sum_{i=1}^{n_s} b_i K_D(s_i, s) \,,$$

where the radii of $\mathcal{H}$ and $\mathcal{D}$ are implicitly given by λ_h and λ_d, respectively.

We note that the computation required by this algorithm is dominated by solving systems of linear equations of size $n_s \times n_s$ and $n_s + 1 \times n_s + 1$, so therefore its computational complexity depends on n_s. Finally, since in practice β is unknown, we can estimate it using $\hat{\varphi}$.

5.2 Solving for Optimal Weights

We now describe a method for analytically computing $\arg\min_W \sup_{g \in \mathcal{G}} \hat{J}_\lambda(W, g)$ for kernel-based classes $\mathcal{G} = \mathcal{G}_{K,\lambda}$, as defined in lemma 1. Our approach is based on the following theorem.

Theorem 7. *Suppose we observe n_z distinct Z tuples in our dataset, which we denote $\{z_1, \ldots, z_{n_z}\}$, and define the index function $\nu : [n] \to [n_z]$ such that $Z_i = z_{\nu(i)}$. In addition, for each $i \in [n_z]$, let U_i and $\tilde{U}_i$ denote iid random variables distributed according to $\hat{\varphi}(z_i)$, let $N \in \mathbb{Z}^{n_z}$ denote the vector of counts of the tuples $\{z_1, \ldots, z_{n_z}\}$ in our dataset, let $W^* = \arg\min_{W \in \mathcal{W}} \sup_{g \in \mathcal{G}} \hat{J}_\lambda(W, g)$ for some given set $\mathcal{W}$, and define*

$$\begin{aligned}
k_{i,j} &= K((s_i, U_i), (s_j, \tilde{U}_j)) \quad \forall i, j \in [n_z] \\
G_{ij} &= N_i N_j \delta_{a_i a_j}\mathbb{E}[k_{i,j}] + \lambda N_i \delta_{ij} \quad \forall i, j \in [n_z] \\
g_i &= N_i \sum_{j=1}^{n_z} \hat{d}(s_j)\mathbb{E}[\pi_e(a_i \mid s_j, \tilde{U}_j)k_{i,j}] \quad \forall i \in [n_z] \\
W' &= \arg\min_{W \in \mathcal{W}} W^T G W - 2g^T W \,.
\end{aligned}$$

Then, $W_i^* = W'_{\nu(i)} \; \forall i \in [n]$.

If we do not constrain the weights W, as in both our theory and our experiments, then

$$W_i^* = (G^{-1}g)_{\nu(i)} \,.$$

That is, we can compute the optimal weights be solving a linear system of equations of size $n_z \times n_z$. Hence, the computational complexity of our optimal balancing algorithm depends on the number of unique Z values in the dataset.

Finally, we note that if we wished to impose some constraints on W, such as $W \in \Delta^n$ (the set of categorical distributions over n categories), then we could instead solve a quadratic program. However, our theory does not support this, and in practice in our experiments we calculate W^* using the unconstrained analytic solution.

6 Experiments

We now evaluate our proposed method and demonstrate its benefits over state-of-the-art baselines for OPE. Our method requires as input an approximate confounder model for the posterior of U given Z: $\hat{\varphi}(z) \approx P(\cdot \mid Z = z)$. Since our baselines cannot account for unmeasured confounding, for fairness we allow these methods access to $\hat{\varphi}$. Specifically, for each $i \in [n]$ we sample a value $\hat{U}_i$ from the approximate posterior $\hat{\varphi}(Z_i)$, and augment the baselines' data with

[7] This is in contrast to Bennett et al. (2019), who used neural networks and smooth-game optimization techniques for their instrumental variable regression estimator.

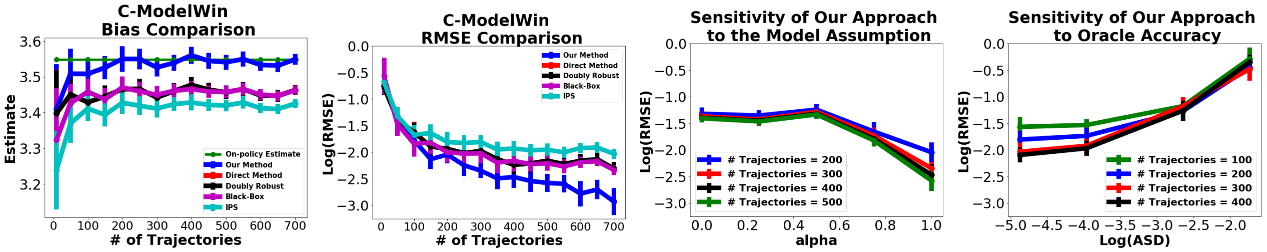

Figure 3: C-ModelWin Results. From left to right: The off-policy estimate, The $\log(\mathrm{RMSE})$ of different methods as we change the number of trajectories, sensitivity of our estimator to model misspecification, and to noise in the confounders posterior distribution.

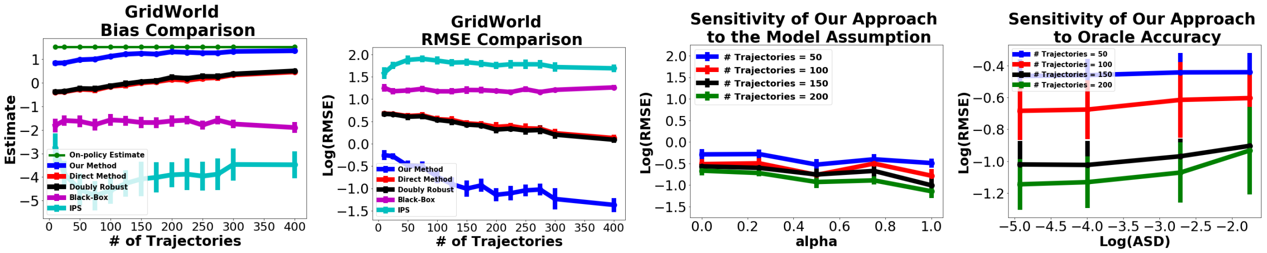

Figure 5: Confounded GridWorld Results. From left to right: off-policy estimate, $\log(\mathrm{RMSE})$ of different methods as we change the number of trajectories, sensitivity of our estimator to model misspecification, and to noise in the confounders posterior distribution. Note that we changed the y axis scale in the last plot for clarity, since the effect was very small.

$\{\hat{U}_i\}_{i\in[n]}$. They then use $(S_i, \hat{U}_i)$ as the state variable rather than S_i. However, since we only assumed a latent variable model for (Z, U), but not for (Z, U, R) (that is, we do not assume an outcome model) we expect that this may still lead to biased estimators even if $\hat{\varphi}$ is perfect.[8]

We consider the following baselines: **Direct Method** which fits an outcome model using the imputed confounders; **Doubly Robust** which combines our optimal balancing weights with the Direct Method, by re-weighting the estimated reward residuals; **Inverse Propensity Scores (IPS)** with IPS weights calculated as in Liu et al. (2018); and **Black-Box** which is state-of-the-art weighted estimator (Mousavi et al., 2020). For a detailed description of these baselines see appendix F.1, and for additional details on hyperparameters for our method and baselines see appendix F.2.

First Experiment. In this experiment we consider the C-ModelWin environment, which is a confounded variant of ModelWin (Thomas & Brunskill, 2016). This is a simple tabular environment with 3 states, 2 actions, and 2 confounder levels. We depict this environment in fig. 2, and describe it in detail in appendix F.3.

First, we compare our estimator $\hat{\tau}_W$, with $\hat{d}$ calculated as in section 5.1 and W estimated as in section 5.2, against the baselines. For this comparison we use the true conditional confounder distribution for $\hat{\varphi}$, with datasets of varying number of trajectories of length 100, and performing 50 repetitions for each configuration of estimator and number of trajectories to compute 95% confidence intervals.[9] We display the results of this comparison in the first two plots of fig. 3, where we plot the estimated policy value and corresponding root mean squared error (RMSE) respectively for every configuration. We see that our estimator achieves strong results, with near-zero bias as we increase the number of trajectories. This is in contrast to the baselines, all of which converge to biased estimates as we increase the number of trajectories, with significantly higher RMSE.

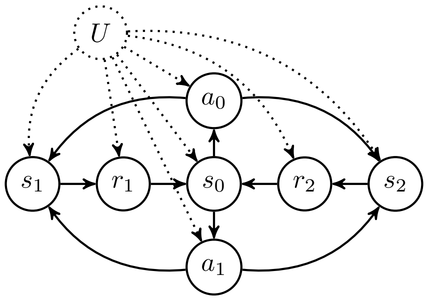

Figure 2: The C-ModelWin environment with 3 states and 2 actions and a confounder.

Next, we investigate the sensitivity of our estimator to the

[8] This is because we can only sample confounders conditioned on Z, not on (Z, R), so the dataset augmented with imputed confounders will be distributed differently to a dataset augmented with the true confounders.

[9] We also used these trajectory lengths and numbers of repetitions in our sensitivity experiments.

assumption of iid confounders. Let $\mathcal{P}_{\text{iid}}$ denote the iid confounder distribution under the C-ModelWin environment, and $\mathcal{P}_{\text{alt}}$ denote some alternative distribution, where within each trajectory the distribution of the confounder at time t depends on the confounder at time $t-1$. We experiment with a variation of C-ModelWin, where confounders are distributed according to $\alpha\mathcal{P}_{\text{iid}} + (1-\alpha)\mathcal{P}_{\text{alt}}$, for some $\alpha \in [0,1]$. Thus, we recover C-ModelWin with $\alpha = 1$, and as we decrease α the iid confounder assumption becomes increasingly violated. The specific alternative model $\mathcal{P}_{\text{alt}}$ used is described in appendix F.4. We display the RMSE of our estimator for various numbers of trajectories and various values of α in the third plot in fig. 3. We see here that, as predicted in section 4.2, the effects of this assumption violation are continuous; when α is close to 1 the RMSE only increases slightly.

Thirdly, we investigate how error in $\hat{\varphi}$ affects our algorithm. We inject error by adding random Gaussian noise of varying variance to the logits of the conditional confounder distribution for each level of Z (before re-normalizing) and measured the amount of noise via the average standard devaition (ASD) metric, which calculates the expected standard deviation of $\hat{P}(U = u \mid Z)$, averaged over the levels of U.[10] Details of this metric and our noise injection are in appendix F.5. We display the RMSE of our estimator under varying levels of noise injection in the fourth plot of fig. 3. We observe that again, as predicted in section 4.2, the effects of noise injection are continuous; as we increase the level of noise injection (as measured by ASD) the RMSE gradually increases, with minimal impact when the error in $\hat{\varphi}$ is very small. We provide additional plots in appendix F.6, repeating both sensitivity experiments for the baselines.

Second Experiment. In this experiment we consider a confounded version of the GridWorld environment. This environment consists of a 10×10 grid, with 4 actions corresponding to attempted movement in each direction, reward based on moving toward the goal, and 2 confounder levels. We depict this envirnment in fig. 4, and describe it in detail in appendix F.3.

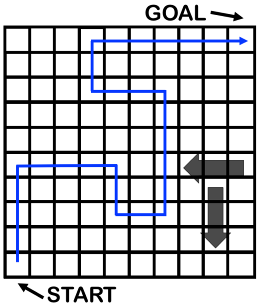

Figure 4: The GridWorld environment.

We conduct the same set of experiments with GridWorld as with C-ModelWin, except that each trajectory was of length 200. We detail the alternative non-iid confounder model used in the sensitivity part of the experiment in appendix F.4, we display the corresponding plots in fig. 5, and we include additional sensitivity results for baselines in appendix F.6. In general, our results here follow the same trend as in the previous experiment. We note that with GridWorld, which is more complex than C-ModelWin, the benefits of our methodology are even more evident, with a larger relative decrease in RMSE compared to baselines. Interestingly, in this setting we see that our method seems especially robust to model assumption violations and nuisance error, with relatively small increases in RMSE in the second two plots. We hypothesize that this is because the setting is more challenging than C-ModelWin, so the error introduced by these perturbations is relatively small compared with the overall errors of the estimators. This suggests that our estimator may be relatively robust to these issues in challenging real-world settings where RMSE is naturally relatively high.

7 Conclusion

In this work, we considered OPE in infinite-horizon reinforcement learning with unobserved confounders. We proposed a novel estimator, and showed its consistency under reasonable assumptions that account for nuisance estimation error and model misspecification. This is in contrast to existing estimators designed for fully-observable MDPs, which typically are unbiased and inconsistent. We also provided sensitivity results bounding the asymptotic bias of our estimator under small violations of these assumptions. Furthermore, we validated our method empirically, demonstrating its accuracy against baselines and corroborating our theoretical analysis.

These promising results open up a number of interesting research directions. First, as an alternative or complement to our optimal balancing-based approach, one could investigate direct or "model-based" approaches, by directly estimating μ, and using this in estimators. This could be combined with our approach to improve accuracy, using the doubly robust augmentation (Kallus & Uehara, 2019a; Tang et al., 2020). Second, we could extend this kind of approach to episodic settings, where there is a fixed time-horizon, by estimating the time-dependent state-density ratio at each time step. Third, we may be able to avoid the dependency on the knowledge of behavior policy by using black-box or behavior-agnostic methods (e.g., Nachum et al., 2019a; Mousavi et al., 2020). Last but not least, one could apply our approach to extend work on batch policy optimization (e.g., Nachum et al., 2019b) to the case of unmeasured confounders.

Acknowledgments. This material is based upon work supported by the National Science Foundation under Grant No. 1846210. This work was done when AM and LL were with Google Research.

[10] With expectation taken over Z, and standard deviation over random noise injection.

References

Kamyar Azizzadenesheli, Alessandro Lazaric, and Animashree Anandkumar. Reinforcement learning of POMDPs using spectral methods. In *Proceedings of the 29th Conference on Learning Theory*, pp. 193–256, 2016.

Elias Bareinboim, Andrew Forney, and Judea Pearl. Bandits with unobserved confounders: A causal approach. *Advances in Neural Information Processing Systems*, 28: 1342–1350, 2015.

Andrew Bennett and Nathan Kallus. Policy evaluation with latent confounders via optimal balance. In *Advances in Neural Information Processing Systems*, pp. 4827–4837, 2019.

Andrew Bennett and Nathan Kallus. The variational method of moments. *arXiv preprint arXiv:2012.09422*, 2020.

Andrew Bennett, Nathan Kallus, and Tobias Schnabel. Deep generalized method of moments for instrumental variable analysis. In *Advances in Neural Information Processing Systems*, pp. 3559–3569, 2019.

Lars Buesing, Theophane Weber, Yori Zwols, Nicolas Heess, Sébastien Racanière, Arthur Guez, and Jean-Baptiste Lespiau. Woulda, coulda, shoulda: Counterfactually-guided policy search. In *Proceedings of the 7th International Conference on Learning Representations (ICLR)*, 2019.

Zhihong Cai and Manabu Kuroki. On identifying total effects in the presence of latent variables and selection bias. In *Proc. of the 24th Conference on Uncertainty in Artificial Intelligence*, pp. 62–69, 2008.

Marine Carrasco, Jean-Pierre Florens, and Eric Renault. Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. *Handbook of econometrics*, 6:5633–5751, 2007.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, and Whitney K Newey. Double machine learning for treatment and causal parameters. Technical report, cemmap working paper, 2016.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bulletin of the American mathematical society*, 39(1):1–49, 2002.

Bo Dai, Ofir Nachum, Yinlam Chow, Lihong Li, Csaba Szepesvári, and Dale Schuurmans. CoinDICE: Off-policy confidence interval estimation. In *Advances in Neural Information Processing Systems 33*, pp. 9398–9411, 2020.

Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. *arXiv preprint arXiv:2006.07201*, 2020.

Jessie K Edwards, Stephen R Cole, and Daniel Westreich. All your data are always missing: incorporating bias due to measurement error into the potential outcomes framework. *International journal of epidemiology*, 44(4): 1452–1459, 2015.

Peter A Frost. Proxy variables and specification bias. *The review of economics and Statistics*, pp. 323–325, 1979.

Carles Gelada and Marc G. Bellemare. Off-policy deep reinforcement learning by bootstrapping the covariate shift. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, pp. 3647–3655, 2019.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, pp. 1029–1054, 1982.

Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. In *Conference on Learning Theory (COLT) Proceedings*, 2009.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.

Nathan Kallus. Generalized optimal matching methods for causal inference. *arXiv preprint arXiv:1612.08321*, 2016.

Nathan Kallus. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, pp. 8895–8906, 2018.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes, 2019a. arXiv:1908.08526.

Nathan Kallus and Masatoshi Uehara. Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning, 2019b. arXiv:1909.05850.

Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning, 2020. arXiv:2002.04518.

Nathan Kallus, Xiaojie Mao, and Madeleine Udell. Causal inference with noisy and missing covariates via matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6921–6932, 2018.

Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

Manabu Kuroki and Judea Pearl. Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2): 423–437, 2014.

Erich L Lehmann and George Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pp. 5356–5366, 2018.

Yao Liu, Pierre-Luc Bacon, and Emma Brunskill. Understanding the curse of horizon in off-policy evaluation via conditional importance sampling. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 6184–6193, 2020.

Christos Louizos, Uri Shalit, Joris M Mooij, David Sontag, Richard Zemel, and Max Welling. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, pp. 6446–6456, 2017.

Chaochao Lu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Deconfounding reinforcement learning in observational settings, 2018. arXiv:1812.10576.

Shahar Mendelson. On the performance of kernel classes. *Journal of Machine Learning Research*, 4(Oct):759–771, 2003.

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

Ali Mousavi, Lihong Li, Qiang Liu, and Denny Zhou. Black-box off-policy estimation for infinite-horizon reinforcement learning. In *Proceedings of the Eighth International Conference on Learning Representations (ICLR)*, 2020.

Krikamol Muandet, Arash Mehrjou, Si Kai Lee, and Anant Raj. Dual iv: A single stage instrumental variable regression. *arXiv preprint arXiv:1910.12358*, 2019.

Susan A. Murphy, Mark van der Laan, and James M. Robins. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, 2001.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-agnostic estimation of discounted stationary distribution corrections. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019a.

Ofir Nachum, Bo Dai, Ilya Kostrikov, Yinlam Chow, Lihong Li, and Dale Schuurmans. Algae: Policy gradient from arbitrary experience, 2019b. CoRR abs/1912.02074.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *arXiv preprint arXiv:2003.05623*, 2020.

Judea Pearl. On measurement bias in causal inference. *arXiv preprint arXiv:1203.3504*, 2012.

Emmanuel Rio. Inequalities and limit theorems for weakly dependent sequences. Lecture, September 2013.

Amirreza Shaban, Mehrdad Farajtabar, Bo Xie, Le Song, and Byron Boots. Learning latent variable models by improving spectral solutions with exterior point method. In *UAI*, pp. 792–801, 2015.

Matthijs T. J. Spaan. Partially observable Markov decision processes. In Marco Wiering and Martijn van Otterlo (eds.), *Reinforcement Learning: State of the Art*, pp. 387–414. Springer Verlag, 2012.

Ziyang Tang, Yihao Feng, Lihong Li, Dengyong Zhou, and Qiang Liu. Doubly robust bias reduction in infinite horizon off-policy estimation. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*, 2020.

Guy Tennenholtz, Shie Mannor, and Uri Shalit. Off-policy evaluation in partially observable environments. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, pp. 2139–2148, 2016.

Masatoshi Uehara, Jiawei Huang, and Nan Jiang. Minimax weight and Q-function learning for off-policy evaluation, 2020.

Aad W Van der Vaart. *Asymptotic statistics*. Cambridge University Press, 2000.

Michael R Wickens. A note on the use of proxy variables. *Econometrica: Journal of the Econometric Society*, pp. 759–761, 1972.

Jeffrey M Wooldridge. On estimating firm-level production functions using proxy variables to control for unobservables. *Economics Letters*, 104(3):112–114, 2009.

Junzhe Zhang and Elias Bareinboim. Markov decision processes with unobserved confounders: A causal approach. Technical Report R-23, Columbia CausalAI Laboratory, 2016.