
Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

# Proximal Reinforcement Learning: Efficient Off-Policy Evaluation in Partially Observed Markov Decision Processes

Andrew Bennett, Nathan Kallus

**To cite this article:**
Andrew Bennett, Nathan Kallus (2023) Proximal Reinforcement Learning: Efficient Off-Policy Evaluation in Partially Observed Markov Decision Processes. Operations Research

Published online in Articles in Advance 26 Sep 2023

. https://doi.org/10.1287/opre.2021.0781

https://pubsonline.informs.org/journal/opre

OPERATIONS RESEARCH
*Articles in Advance*, pp. 1–16
ISSN 0030-364X (print), ISSN 1526-5463 (online)



Crosscutting Areas

# Proximal Reinforcement Learning: Efficient Off-Policy Evaluation in Partially Observed Markov Decision Processes

**Andrew Bennett,**[a] **Nathan Kallus**[a,*]

[a] Cornell Tech, Cornell University, New York, New York 10044
*Corresponding author
**Contact:** awb222@cornell.edu (AB); kallus@cornell.edu, https://orcid.org/0000-0003-1672-0507 (NK)

**Received:** December 14, 2021
**Revised:** March 15, 2023
**Accepted:** July 27, 2023
**Published Online in Articles in Advance:** September 26, 2023

**Area of Review**: Machine Learning and Data Science

**https://doi.org/10.1287/opre.2021.0781**



**Abstract.** In applications of offline reinforcement learning to observational data, such as in healthcare or education, a general concern is that observed actions might be affected by unobserved factors, inducing confounding and biasing estimates derived under the assumption of a perfect Markov decision process (MDP) model. Here we tackle this by considering off-policy evaluation in a partially observed MDP (POMDP). Specifically, we consider estimating the value of a given target policy in an unknown POMDP given observations of trajectories with only partial state observations and generated by a different and unknown policy that may depend on the unobserved state. We tackle two questions: what conditions allow us to identify the target policy value from the observed data and, given identification, how to best estimate it. To answer these, we extend the framework of proximal causal inference to our POMDP setting, providing a variety of settings where identification is made possible by the existence of so-called bridge functions. We term the resulting framework proximal reinforcement learning (PRL). We then show how to construct estimators in these settings and prove they are semiparametrically efficient. We demonstrate the benefits of PRL in an extensive simulation study and on the problem of sepsis management.

**Funding:** This work was supported by the National Science Foundation [Grant 1846210].
**Supplemental Material:** The online appendix is available at https://doi.org/10.1287/opre.2021.0781.

**Keywords:** offline reinforcement learning • unmeasured confounding • semiparametric efficiency

## 1. Introduction

An important problem in reinforcement learning (RL) is off-policy evaluation (OPE), which is defined as estimating the average reward generated by a target *evaluation* policy, given observations of data generated by running some different *behavior* policy. This problem is particularly important in many application areas such as healthcare, education, or robotics, where experimenting with new policies may be expensive, impractical, or unethical. In such applications OPE may be used to estimate the benefit of proposed policy changes by decision makers or as a building block for the related problem of policy optimization. At the same time, in the same applications, unobservables can make this task difficult due to the lack of experimentation.

As an example, consider the problem of evaluating a newly proposed policy for assigning personalized curricula to students semester by semester, where the curriculum assignment each semester is decided based on observed student covariates, such as course outcomes and aptitude tests, with the goal of maximizing student outcomes as measured, for example, by standardized test scores. Because it may be unethical to experiment with potentially detrimental curriculum plans, we may wish to evaluate such policies based on passively collected data where the targeted curriculum was decided by teachers. However, there may be factors unobserved in the data that jointly influence the observed student covariates, curriculum assignments, and student outcomes; this may arise for example because the teacher can perceive subjective aspects of the students' personalities or aptitudes and take these into account in their decisions. Although such confounding breaks the usual Markovian assumptions that underlie standard approaches to OPE, the process may well be modeled by a partially observed Markov decision process (POMDP). Two key questions for OPE in POMDPs are: when is policy value still identifiable despite confounding due to partial observation and, when it is, how can we estimate it most efficiently.

In this work, we tackle these two questions, expanding the range of settings that enable identification and providing efficient estimators in these settings. First, we extend an existing identification result for OPE in

tabular POMDPs (Tennenholtz et al. 2020) to the continuous setting, which provides some novel insight on this existing approach but also highlights its limitations. To break these limitations, motivated by these insights, we provide a new general identification result based on extending the proximal causal inference framework (Miao et al. 2018a, Cui et al. 2020, Kallus et al. 2022) to the dynamic, longitudinal setting. This permits identification in more general settings. Unlike the previous results, this one expresses the value of the evaluation policy as the mean of some score function under the distribution over trajectories induced by the logging policy, which allows for natural estimators with good qualities. In particular, we prove appropriate conditions under which the estimators arising from this result are consistent, asymptotically normal, and semiparametrically efficient. In addition, we provide a tractable algorithm for computing the nuisance functions that allow such estimators to be computed, based on recent state-of-the-art methods for solving conditional moment problems. We term this framework proximal reinforcement learning (PRL), highlighting the connection to proximal causal inference. We finally provide a series of experiments, on both a synthetic toy scenario and a complex scenario based on a sepsis simulator, which empirically validate our theoretical results and demonstrate the benefits of PRL.

## 2. Related Work

First, there is an extensive line of recent work on OPE under unmeasured confounding. This work considers many different forms of confounding, including confounding that is independent and identically distributed (i.i.d.) at each time step (Bennett et al. 2021, Liao et al. 2021, Wang et al. 2021), occurs only at a single time step (Namkoong et al. 2020), satisfies a "memorylessness" property (Kallus and Zhou 2020), follows a POMDP structure (Oberst and Sontag 2019, Tennenholtz et al. 2020, Nair and Jiang 2021, Killian et al. 2022), may take an arbitrary form (Chandak et al. 2021, Chen and Zhang 2023), or is in fact not a confounder (Hu and Wager 2023). These works have varying foci: Namkoong et al. (2020), Kallus and Zhou (2020), and Chen and Zhang (2023) focus on computing intervals comprising the partial identification set of all hypothetical policy values consistent with the data and their assumptions; Oberst and Sontag (2019) and Killian et al. (2022) focus on sampling counterfactual trajectories under the evaluation policy given that the POMDP follows a particular Gumbel-softmax structure; Wang et al. (2021) and Gasse et al. (2021) focus on using the offline data to warm start online reinforcement learning; Liao et al. (2021) study OPE using instrumental variables; Chandak et al. (2021) show that OPE can be performed under very general confounding if the behavior policy probabilities of the logged actions are known; Hu and Wager (2023) consider hidden states that do not affect the behavior policy and are therefore not confounders but do make OPE harder by breaking Markovianity thereby inducing a curse of horizon; and Tennenholtz et al. (2020) and Nair and Jiang (2021) study conditions under which the policy value under the POMDP model is identified.

Of the past work on OPE under unmeasured confounding, Tennenholtz et al. (2020) and Nair and Jiang (2021) are closest to ours because they too consider a general POMDP model of confounding, namely without restrictions that preserve Markovianity via i.i.d. confounders, knowing the confounder-dependent propensities, having unconfounded logged actions, or using a specific Gumbel-softmax form. Tennenholtz et al. (2020) consider a particular class of tabular POMDPs satisfying some rank constraints, and Nair and Jiang (2021) extend these results and slightly relax its assumptions. However, neither considers how to actually construct OPE estimators based on their identification results that satisfy desirable properties such as consistency or asymptotic normality, and they can only be applied to tabular POMDPs. Our work presents a novel and general identification result and proposes a class of resulting OPE estimators that possesses such desirable properties.

Another area of relevant literature is on proximal causal inference (PCI). PCI was first proposed by Miao et al. (2018a), showing that using two conditionally independent proxies of the confounder (known as a negative control outcome and a negative control action), we can learn an outcome bridge function that generalizes the standard mean-outcome function and controls for the confounding effects. Since then, this work has been expanded, including by alternatively using an action bridge function that instead generalizes the inverse propensity score (Miao et al. 2018b), allowing for multiple fixed treatments (Tchetgen Tchetgen et al. 2020), performing multiply-robust treatment effect estimation (Shi et al. 2020), combining outcome and action bridge functions for semiparametrically efficient estimation (Cui et al. 2020), using PCI to estimate the value of contextual-bandit policies (Xu et al. 2021) or generalized treatment effects (Kallus et al. 2022), or estimating bridge functions using adversarial machine learning (Ghassami et al. 2022, Kallus et al. 2022). In addition, the OPE for POMDP methodologies of Tennenholtz et al. (2020) and Nair and Jiang (2021) discussed earlier were said to be motivated by PCI. Our paper relates to this body of work as it proposes a new way of performing OPE for POMDPs using PCI, and it also proposes a new adversarial machine learning-based approach for estimating the bridge functions.

At the intersection of work of OPE and PCI is the concurrent work of Ying et al. (2021), which considers PCI in multi–time step scenarios, given two proxies at each time step similar to what we consider in Section 4.2.

Unlike us, they only consider the problem of estimating treatment effects for fixed vectors of treatment at each time step, optionally conditional on observable context at $t=1$, as opposed to evaluating policies that can adaptively treat based on the context available thus far.

Finally, there is an extensive body of work on learning policies for POMDPs using online learning. For example, see Azizzadenesheli et al. (2016), Katt et al. (2017), Bhattacharya et al. (2020), Yang et al. (2021), and Singh et al. (2021), and references therein. Our work is distinct in that we consider an offline setting where identification is an issue. At the same time, our work is related to the online setting in that it could potentially be used to augment and warm start such approaches if there is also offline observed data available.

## 3. Problem Setting

A POMDP is formally defined by a tuple $(\mathcal{S},\mathcal{A},\mathcal{O},H,P_O,P_R,P_T)$, where $\mathcal{S}$ denotes a state space, $\mathcal{A}$ denotes a finite action space, $\mathcal{O}$ denotes an observation space, $H\in\mathbb{N}$ denotes a time horizon, $P_O$ is an observation kernel, with $P_O^{(t)}(\cdot|s)$ denoting the density of the observation $O_t$ given the state $S_t=s$ at time $t$, $P_R$ is a reward kernel, with $P_R^{(t)}(\cdot|s,a)$ denoting the density of the (bounded) reward $R_t\in[-R_{\max},R_{\max}]$ given the state $S_t=s$ and action $A_t=a$ at time $t$, and $P_T$ is a transition kernel, with $P_T^{(t)}(\cdot|s,a)$ denoting the density of the next $S_{t+1}$ given the state $S_t=s$ and action $A_t=a$ at time $t$. We allow for the POMDP to be time inhomogeneous; that is, we allow the outcome, reward, and transition kernels to potentially depend on the time index. Finally, we let $O_0$ denote some prior observation of the state before $t=1$ (which may be empty), and we let $\tau_t^{\text{full}}$ and $\tau_t$ denote the true and observed trajectories up to time $t$, respectively, which we define as

$$\begin{aligned}\tau_0 &= \tau_0^{\text{full}} = O_0\\ \tau_t &= (O_0,(O_1,A_1,R_1),(O_2,A_2,R_t),\dots,(O_t,A_t,R_t))\\ \tau_t^{\text{full}} &= (O_0,(S_1,O_1,A_1,R_1),(S_2,O_2,A_2,R_t),\\ &\qquad \dots,(S_t,O_t,A_t,R_t)).\end{aligned}$$

Let $\pi_b$ be some given randomized *logging policy*, which is characterized by a sequence of functions $\pi_b^{(1)},\dots,\pi_b^{(H)}$, where $\pi_b^{(t)}(a|S_t)$ denotes the probability that the logging policy takes action $a\in\mathcal{A}$ at time $t$ given state $S_t$. The logging policy together with the POMDP define a joint distribution over the (true) trajectory $\tau_H^{\text{full}}$ given by acting according to $\pi_b$; let $\mathcal{P}_b$ denote this distribution. All probabilities and expectations in the ensuing will be with respect to $\mathcal{P}_b$ unless otherwise specified, for example, by a subscript.

Our data consist of observed trajectories generated by the logging policy: $\mathcal{D}=\{\tau_H^{(1)},\tau_H^{(2)},\dots,\tau_H^{(n)}\}$, where each $\tau_H^{(i)}$ is an i.i.d. sample of $\tau_H$ (which does not contain $S_t$), distributed according to $\mathcal{P}_b$. Importantly, we emphasize that, although we assume that states are unobserved by the decision maker and are not included in the logged data $\mathcal{D}$, the logging policy still uses these hidden states, inducing confounding.

Implicit in our notation $\pi_b^{(t)}(a|S_t)$ is that the logging policy actions are independent of the past given current state $S_t$. Similarly, the POMDP model is characterized by similar independence assumption with respect to observation and reward emissions, and state transitions. This means that $\mathcal{P}_b$ satisfies a Markovian assumption with respect to $S_t$; however, as $S_t$ is unobserved, we cannot condition on it and break the past from the future. We visualize the directed acyclic graph (DAG) representing $\mathcal{P}_b$ in Figure 1. In particular, we have the following conditional independencies in $\mathcal{P}_b$: For every $t$,

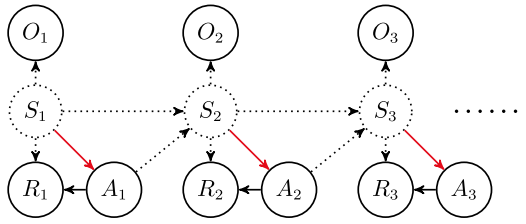
$$\begin{aligned}&O_t \perp\!\!\!\perp \tau_{t-1}^{\text{full}}|S_t, \quad R_t \perp\!\!\!\perp \tau_{t-1}^{\text{full}},O_t|S_t,A_t,\\ &S_{t+1} \perp\!\!\!\perp \tau_{t-1}^{\text{full}},O_t,R_t|S_t,A_t, \quad A_t \perp\!\!\!\perp \tau_{t-1}^{\text{full}}|S_t.\end{aligned}$$

Now, let $\pi_e$ be some deterministic *target policy* that we wish to evaluate, which is characterized by a sequence of functions $\pi_e^{(1)},\dots,\pi_e^{(H)}$, where $\pi_e^{(t)}(O_t,\tau_{t-1})\in\mathcal{A}$ denotes the action taken by policy $\pi_e$ at time $t$ given current observation $O_t$ and the past observable trajectory $\tau_{t-1}$. We visualize the POMDP model under such a policy that only depends on observable data in Figure 2. We allow $\pi_e^{(t)}$ to potentially depend on all observable data up to time $t$; this is because the Markovian assumption *does not* hold with respect to the observations $O_t$, so we may wish to consider policies that use all past observable information to best account for the unobserved state. We let $\mathcal{P}_e$ denote the distribution over trajectories that would be obtained by following policy $\pi_e$ in the POMDP. Then, given some discounting factor $\gamma\in(0,1]$, we define the *value* of policy $\pi_e$ as

$$v_\gamma(\pi_e)=\sum_{t=1}^{H}\gamma^{t-1}\mathbb{E}_{\mathcal{P}_e}[R_t].$$

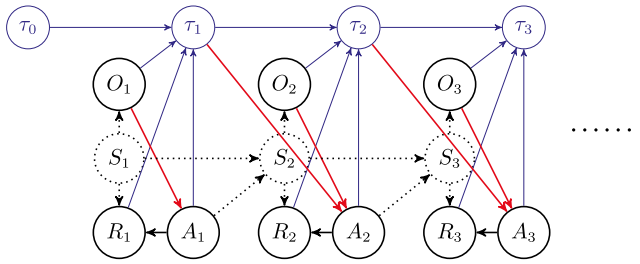
The task OPE under the POMDP model is to estimate $v_\gamma(\pi_e)$ (a function of $\mathcal{P}_e$) given $\mathcal{D}$ (drawn from $\mathcal{P}_b$).

**Figure 1.** (Color online) POMDP Model Under Logging Policy $\pi_b$




*Notes.* The arrows from $S_t$ to $A_t$ (red online) make explicit the dependence of $\pi_b$ on the hidden state. Dashed circles denote variables unobserved in our data.

**Figure 2.** (Color online) POMDP Model Under Evaluation Policy $\pi_e$




*Note.* The arrows from $O_t$ to $A_t$ and from $\tau_t$ to $A_{t+1}$ (red online) make explicit the dependence of $\pi_e$ on the current observation and previous observable trajectory, and the nodes $\tau_t$ and arrows into them (blue online) make explicit the dependence of the observable trajectories on the data.

## 4. Identification Theory

Before considering how to actually estimate $v_\gamma(\pi_e)$, we first consider the problem of *identification*, which is the problem of finding some function $\psi$ such that $v_\gamma(\pi_e) = \psi(\mathcal{P}_b)$ and is a prerequisite for identification. This is the first stepping stone because $\mathcal{P}_b$ is the most we could hope to ever learn from observing $\mathcal{D}$. If such a $\psi$ exists, then we say that $v_\gamma(\pi_e)$ is *identified* with respect to $\mathcal{P}_b$. In general, such an identification result is impossible for the OPE problem given unobserved confounding as introduced by our POMDP model. Therefore, we must impose some assumptions on $\mathcal{P}_b$ for such identification to be possible.

To the best of our knowledge, the only existing identification result of this kind was presented by Tennenholtz et al. (2020) (with a slight generalization given by Nair and Jiang 2021) and is only valid in tabular settings where states and observations are discrete. We will proceed first by extending this approach to more general, nontabular settings. However, we will note that there are some restrictive limitations to estimation based on this approach. Therefore, motivated by the limitations, we develop a new and more general identification theory that extends the PCI approach to the sequential setting and easily enables efficient estimation.

### 4.1. Identification by Time-Independent Sampling and Its Limitations

For our generalization of Tennenholtz et al. (2020), we will consider evaluating policies $\pi_e$ such that $\pi_e^{(t)}(O_t, \tau_t)$ only depends on $O_{1:t}$ and $A_{1:t-1}$; that is, $\pi_e^{(t)}$ can depend on all observed data available at time $t$ except for $O_0$ and past rewards. First, for each $t \in \{1, \dots, H\}$, let $D_t = (O_{t-1}, O_t, O_{t+1}, A_t, R_t)$, and for any such tuple $D = (O, O', O'', A, R)$ define $o(D) = O$, $o'(D) = O'$, $o''(D) = O''$, $a(D) = A$, and $r(D) = R$. In addition, define the shorthand $\pi_e^{(t)}(D_{1:t}) = \pi_e^{(t)}(o'(D_t), \dots, o'(D_1), a(D_{t-1}), \dots, a(D_1))$. Furthermore, let $\mathcal{P}_{\text{ind}}$ denote the measure on $D_{1:H}$ in which each tuple $D_t$ is sampled *independently* according to its marginal distribution in $\mathcal{P}_b$. Under this measure, the overlapping observations between these tuples (e.g., $o'(D_t)$ and $o(D_{t+1})$) may take different values. Then, given these definitions, we have the following result.

**Theorem 1.** *Under some regularity conditions detailed in Section EC.1 of the online appendix, there exist functions $\rho^{(t)}$ defined by conditional moment restrictions under $\mathcal{P}_b$, such that for every $t \in \{1, \dots, H\}$, we have*

$$\mathbb{E}_{\mathcal{P}_e}[R_t] = \mathbb{E}_{\mathcal{P}_{\text{ind}}}\left[ r(D_t) \prod_{s=1}^{t} \mathbb{1}\{a(D_s) = \pi^{(t)}(D_{1:s})\} \times \rho^{(s)}(o(D_s), a(D_s), o''(D_{s-1})) \right]. \quad (1)$$

*Furthermore, under the conditions of Tennenholtz et al. (2020, theorem 1), these regularity conditions are satisfied, and the right-hand side (RHS) of Equation (1) is identical to their identification quantity.*

Because $\mathcal{P}_{\text{ind}}$ is a function of $\mathcal{P}_b$, the RHS of Equation (1) is a valid identification quantity, and applying this result for each $t \in [H]$ identifies $v_\gamma(\pi_e)$. The full details of the regularity conditions and nuisance functions governing this result are not very important for this work, so they are deferred along with the proof of this theorem to Section EC.1 of the online appendix. For our purposes, the main takeaway of Theorem 1 is that there exists a natural generalization of Tennenholtz et al. (2020, theorem 1) to nondiscrete settings; although that result was originally expressed as a sum over all possible observable trajectories, we show that it can instead be expressed as the expectation of a simple, estimable quantity whose existence does not depend on discreteness. Unfortunately, the expectation that naturally arises is under $\mathcal{P}_{\text{ind}}$ rather than $\mathcal{P}_b$. This means that empirical approximations of this expectation given $n$ i.i.d. samples from $\mathcal{P}_b$ would require averaging over $n^s$ terms, introducing a curse of dimension. Furthermore, this expectation clearly does not have many of the desirable properties for OPE estimating equations held by many OPE estimators in the simpler MDP setting, such as Neyman orthogonality (Kallus and Uehara 2020, 2022).

### 4.2. Identification by PCI

We now discuss an alternative way of obtaining identifiability, via a reduction to a nested sequence of PCI problems of the kind described by Cui et al. (2020). These authors considered identifying the average treatment effect (ATE) and other related causal estimands for binary decision making problems with unmeasured confounding given two independent proxies for the confounders, one of which is conditionally independent from treatments given confounders, and the other of which is independent from outcomes given treatment and confounders. We will in fact leverage the refinement

of the PCI approach by Kallus et al. (2022), which has strictly weaker assumptions than Cui et al. (2020).

Our reduction works by defining random variables $Z_t$ and $W_t$ for each $t \in [H]$ that are measurable with respect to (w.r.t.) the observed trajectory $\tau_H$, as well as defining random variables $U_t$ for each $t \in [H]$ such that $S_t$ is measurable w.r.t. $U_t$. We, respectively, refer to $Z_t$ and $W_t$ as *negative control actions* and *negative control outcomes*, and we refer to $U_t$ as *confounders*. All triplets $(Z_t, W_t, U_t)$ must be satisfy certain independence properties outlined below. Any definition of such variables that satisfy these independence properties is considered a valid PCI reduction, and we will have various examples of valid PCI reductions for our POMDP model at the end of this section.

To formalize these assumptions, we must first define some additional notation. Let $\mathcal{P}_t^*$ denote the measure on trajectories induced by running policy $\pi_e$ for the first $t-1$ actions and running policy $\pi_b$ henceforth. According to this definition, $\mathcal{P}_b = \mathcal{P}_1^*$, and $\mathcal{P}_e = \mathcal{P}_{H+1}^*$. In addition, let $\mathbb{E}_t^*$ and $P_t^*$ be shorthand for expectation and probability mass under $\mathcal{P}_t^*$, respectively. We visualize these intervention distributions in the first part of Figure 3.

Next, for each $t \in \{1, \dots, H\}$ we define $E_t = \pi_e^{(t)}(O_t, \tau_{t-1})$, and $D_t = (Z_t, W_t, A_t, E_t, R_t)$. In addition, we will refer to any random variable $Y_t$ as an *outcome variable at time t* if it is measurable w.r.t. $(R_t, D_{t+1:H})$. For any such variable and $a \in \mathcal{A}$, we use $Y_t(a)$ to denote a random variable with the same distribution that $Y_t$ would have if, possibly counter to fact, action $a$ were taken at time $t$ instead of $A_t$. Under $\mathcal{P}_t^*$, we can interpret $Y_t(a)$ as the outcome that would be obtained by applying $\pi_e$ for the first $t-1$ actions, the fixed action $a$ at time $t$, and then $\pi_b$ henceforth (as opposed to the factual outcome $Y_t$ obtained by applying $\pi_e$ for the first $t-1$ actions and $\pi_b$ henceforth). According to this notation, $Y_t(A_t) = Y_t$ always.

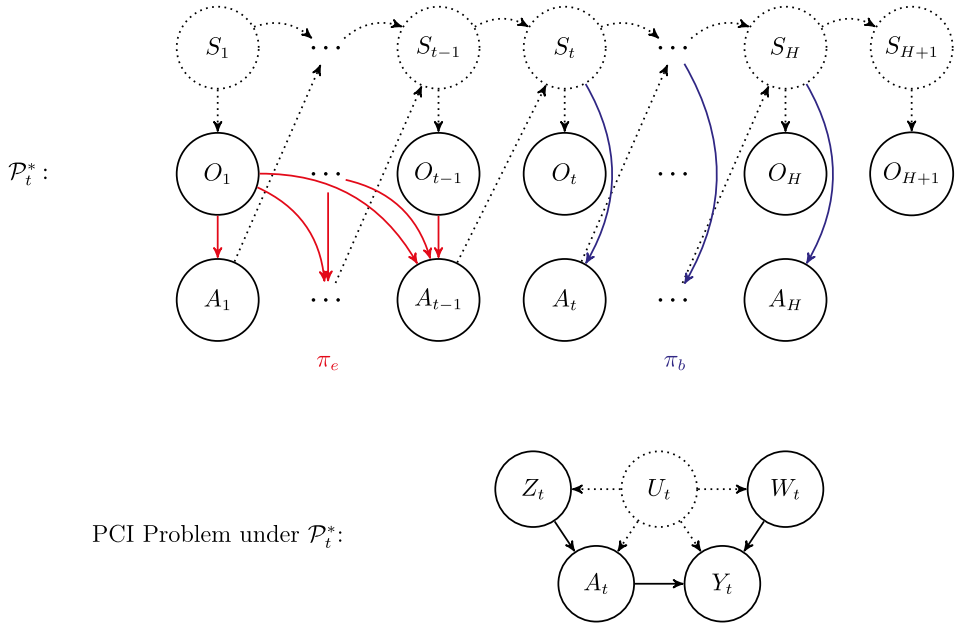
Given these definitions, we are ready to present our core assumptions. Our first assumption is that the confounders $U_t$ are sufficient to induce a particular conditional independence structure between the proxies $Z_t$ and $W_t$, as well as the observable data. Specifically, we assume the following.

**Assumption 1** (Negative Controls). *For each $t \in [H]$ and $a \in \mathcal{A}$, and any outcome variable $Y_t$ that is measurable w.r.t. $(R_t, D_{t+1:H})$, we have*

$$Z_t, A_t \perp\!\!\!\perp_{\mathcal{P}_t^*} W_t, E_t, Y_t(a) \mid U_t.$$

These independence assumptions imply that the decision-making problem under $\mathcal{P}_t^*$ with confounder $U_t$, negative controls $Z_t$ and $W_t$, action $A_t$, and outcome $(R_t, D_{t+1:H})$ satisfy the PCI problem structure as in Cui et al. (2020). We visualize this structure for the problem at time $t$ in Figure 3. In addition, it requires that the action-side proxy $Z_t$ is conditionally independent from the next action $E_t$ that would have been taken under $\pi_e^{(t)}$. We may additionally include an observable context variable $X_t$, which may be useful for defining more efficient

**Figure 3.** (Color online) The Interventional Distribution $\mathcal{P}_t^*$ and the PCI Problem Under It




*Notes.* (Top) Visual representation of the interventional distribution $\mathcal{P}_t^*$. This is the distribution over trajectories obtained by taking actions following the target policy $\pi_e$ for the first $t-1$ actions and then taking all subsequent actions following $\pi_b$. (Bottom) Probabilistic graphical representation of the corresponding proximal causal inference decision-making problem at time $t$ under $\mathcal{P}_t^*$, with outcome variable $Y_t = \phi(R_t, D_{t+1:H})$ for arbitrary $\phi$. The variables $Z_t$ and $W_t$ are conditionally independent action-side and outcome-side proxies for the true (unobserved) confounder $U_t$.

reductions. In this case, the conditional independence assumption in Assumption 1 should hold given both $U_t$ and $X_t$, and in everything that follows $Z_t$, $W_t$, and $U_t$ should be replaced with $(Z_t, X_t)$, $(W_t, X_t)$, and $(U_t, X_t)$, respectively, as in Cui et al. (2020). However, we omit $X_t$ from the notation in the rest of the paper for brevity.

Next, our results require the existence of some *bridge* functions, as follows.

**Assumption 2** (Bridge Functions Exist). *For each $t \in [H]$ and $a \in \mathcal{A}$, and any given outcome variable $Y_t = \phi(R_t, D_{t+1:H})$, there exists functions $q^{(t)}$ and $h^{(t,\phi)}$ satisfying*

$$\mathbb{E}_t^*[q^{(t)}(Z_t, A_t) \mid U_t, A_t = a] = P_t^*(A_t = a \mid U_t)^{-1} \quad a.s.$$

*and* $\mathbb{E}_t^*[h^{(t,\phi)}(W_t, A_t) \mid U_t, A_t = a]$

$$= \mathbb{E}_t^*[\mathbb{1}\{E_t = A_t\} Y_t \mid U_t, A_t = a] \quad a.s..$$

Implicit in the assumption is that $P_t^*(A_t = a \mid U_t) > 0$. We refer to the functions $q^{(t)}$ as *action bridge functions* and $h^{(t,\phi)}$ as *outcome bridge functions*. These may be seen as analogues of inverse propensity scores and state-action quality functions, respectively. As argued previously by Kallus et al. (2022), assuming the existence of these functions is more general than the approach taken by Cui et al. (2020), who require complex completeness conditions. We refer readers to Kallus et al. (2022) for a detailed presentation of conditions under which the existence of such bridge functions can be justified, as well as concrete examples of bridge functions when the negative controls are discrete, or the negative controls and $Y_t$ are defined by linear models.

In the case of both Assumptions 1 and 2, the assumption depends on the choice of proxies $Z_t$ and $W_t$ and on the choice of confounders $U_t$. In addition, the parts of $(O_t, \tau_{t-1})$ that $\pi_e^{(t)}$ may depend on determines what variables $E_t$ is a function of, so the evaluation policy $\pi_e$ also affects the validity of Assumption 1. For now, we just emphasize this important point and present our main identification theory, which is valid given these assumptions. However, we will provide some concrete examples of feasible and valid choices of $(Z_t, W_t, U_t)$ that satisfy Assumption 1 for different kinds of policies $\pi_e$ in Section 4.3. In addition, we provide an in-depth examination of the additional conditions under which Assumption 2 holds for an example tabular setting in Section 4.4.

**Theorem 2.** *Let Assumptions 1 and 2 hold. Define $q^{(t)}$ and $h^{(t)}$ as any solutions to*

$$\mathbb{E}_t^*[q^{(t)}(Z_t, A_t) \mid W_t, A_t = a] = P_t^*(A_t = a \mid W_t)^{-1} \quad a.s. \quad \forall a \in \mathcal{A}, \tag{2}$$

$$\mathbb{E}_t^*[h^{(t)}(W_t, A_t) \mid Z_t, A_t = a] = \mathbb{E}_t^*[\mathbb{1}\{E_t = A_t\} Y_t \mid Z_t, A_t = a] \quad a.s. \quad \forall a \in \mathcal{A}, \tag{3}$$

*where $Y_H = R_H$, and for every $t \le H$, we recursively define*

$$Y_{t-1} = R_{t-1} + \gamma \left( \sum_{a \in \mathcal{A}} h^{(t)}(W_t, a) + q^{(t)}(Z_t, A_t) \times (\mathbb{1}\{A_t = E_t\} Y_t - h^{(t)}(W_t, A_t)) \right). \tag{4}$$

*Also, let $\eta_t = \prod_{s=1}^{t-1} \mathbb{1}\{E_s = A_s\} q^{(s)}(Z_s, A_s)$. Then, we have $v_\gamma(\pi_e) = \mathbb{E}_{\mathcal{P}_b}[\psi_{\mathrm{DR}}(\tau_H)]$, where*

$$\psi_{\mathrm{DR}}(\tau_H) = \sum_{t=1}^{H} \gamma^{t-1} \left( \eta_{t+1} R_t + \eta_t \sum_{a \in \mathcal{A}} h^{(t)}(W_t, a) - \eta_t q^{(t)}(Z_t, A_t) h^{(t)}(W_t, A_t) \right). \tag{5}$$

Because $\mathbb{E}_{\mathcal{P}_b}[\psi_{\mathrm{DR}}(\tau_H)]$ is fully defined by $\mathcal{P}_b$, this is a valid identification result. As detailed in our proof, the existence of solutions to Equations (2) and (3) is guaranteed given our assumptions. Comparing with Theorem 1, this result has many immediate advantages; it is written as an expectation over $\mathcal{P}_b$ and therefore may be analyzed readily using standard semiparametric efficiency theory, and although Equations (2) and (3) may appear complex given that they are expressed in terms of the intervention distributions $\mathcal{P}_t^*$, this can easily be dealt with as discussed later. We also observe that Equation (5) has a very similar structure to the double reinforcement learning (DRL) estimators for the MDP setting (Kallus and Uehara 2020), where $h^{(t)}$ and $q^{(t)}$ are used in place of inverse propensity score and quality function terms, respectively. This is very promising because DRL estimators enjoy desirable properties such as semiparametric efficiency in the MDP setting (Kallus and Uehara 2020). Indeed, in Section 5, we show that similar properties extend to estimators defined based on Equation (5).

At a high level, the proof of Theorem 2 works by defining a series of of outcome variables $Y_t$ such that, for each PCI problem at time $t \in [H]$ under distribution $\mathcal{P}_t^*$ and with outcome variable $Y_t$, the policy value obtained by intervening at time $t$ with $\pi_e$ is equal to $\mathbb{E}_{\mathcal{P}_e}[R_t + \gamma R_{t+1} + \cdots + \gamma^{H-t} R_{H}]$. In the base case of $t = H$, this property is trivially satisfied with $Y_t = R_t$, because under $\mathcal{P}_H^*$, all prior actions prior to time $H$ are taken following $\pi_e$. Conversely, for $t < H$, we establish via backward induction that this holds with $Y_t$ defined according to Equation (4); this works because the term multiplied by $\gamma$ in Equation (4) is the doubly robust influence function for the PCI problem at time $t$, so $\mathbb{E}_t^*[Y_{t-1}] = \mathbb{E}_{\mathcal{P}_e}[R_{t-1}] + \gamma \mathbb{E}_{t+1}^*[Y_t]$. Similarly, $\psi_{\mathrm{DR}}(\tau_H)$ is the doubly robust influence function for the PCI problem at $t = 1$ and so $\mathbb{E}_{\mathcal{P}_b}[\psi_{\mathrm{DR}}(\tau_H)] = \mathbb{E}_2^*[Y_1] = \cdots = v_\gamma(\pi_e)$. That is, we recursively apply the improved identification theory of Kallus et al. (2022) to a nested sequence of

PCI problems. In each step of the induction, we apply Assumptions 1 and 2 with the specific outcome variable $Y_t$. We provide full proof details in Section EC.2 of the online appendix, where we also present a slightly more general result that allows for alternatives to $\psi dr$ that instead resemble importance sampling or direct method estimators for the MDP setting.

### 4.3. Specific Proximal Causal Inference Reductions and Resulting Identification

Next, we provide some discussion of how to actually construct a valid PCI reduction; that is, how to choose $Z_t$, $W_t$, and $U_t$ that satisfy Assumption 1. We provide several options of how this reduction may be performed and discuss in each case the assumptions that would be required of the POMDP and $\pi_e$ for identification based on our results. In all cases that we consider, we would need to additionally justify Assumption 2, which implicitly requires some additional completeness conditions on the choices of $Z_t$, $W_t$, and $U_t$. Furthermore, the practicality of any given reduction would depend heavily on how well correlated $W_t$ and $Z_t$ are for each $t$, which in turn would impact how easily the required nuisance functions $q^{(t)}$ and $h^{(t)}$ could be fit. We summarize these reductions in Table 1.

**4.3.1. Current and Previous Observations.** Perhaps the most simple kind of PCI reduction would be to define $U_t = S_t$, $W_t = O_t$, and $Z_t = (O_{t-1}, A_{t-1}, R_{t-1})$. That is, we use the current hidden state as confounders, and we use both the observation of $S_t$ and the previous observation, action, and reward triple as proxies for $O_t$. For this definition, we define $A_0 = R_0 = \emptyset$. It is easy to verify that this is a valid PCI reduction (i.e., satisfying Assumption 1) as long as $\pi_e^{(t)}$ depends on $(\tau_t, O_t)$ via $O_t$ only. In addition, it is easy to verify that this reduction remains valid if we replace $Z_t$ with $O_{t-1}$, which gives us a very simple and elegant reduction at the slight cost of fewer treatment-side proxies.

This kind of reduction may be relevant in applications where the current observation of the state is considered to be rich enough for decision making, but where nonetheless it is possible that confounding is present. One example of such a setting is a noisy observation setting, where $O_t$ is a direct observation of $S_t$ that may be corrupted with some probability, as discussed in more detail in Section 6. Another example where such a reduction may be desirable is when we wish to consider policies that are functions of $O_t$ only for reasons of simplicity/interpretability. For example, if we wish to evaluate an automated policy for sepsis management, we may wish that the policy is a simple function of the patient's current state that can be understood and audited by doctors.

**4.3.2. Current and *k*-Prior Observation.** An alternative to the previous reduction would be to define define $U_t = (S_t, S_{t-k'+1})$, $W_t = O_t$, and $Z_t = O_{t-k'}$, for some integer $k \geq 2$, where $k' = \min(k, t)$. In this reduction, we can no longer include any action or reward in $Z_t$, as this would break Assumption 1 in general given the definition of $\mathcal{P}_t^*$. This reduction allows for any policy where $\pi_e^{(t)}$ depends on $(\tau_t, O_t)$ via the data from the $k$-most recent time steps; that is, $(O_{t-k'+1:t}, A_{t-k'+1:t-1}, R_{t-k'+1:t-1})$.

This kind of reduction would be useful in applications where it is necessary to consider policies that consider a past history of observations rather than only the most recent observation. For example, if we were considering the task of training a robot to act within an environment that it can only observe part of at each time step through its camera, it may be necessary to consider policies that use several recent observations to build a more accurate map of the environment. However, one limitation of this reduction compared with the previous is that it uses two states as its confounder, which may make Assumption 2 more difficult to satisfy. In addition, because $Z_t$ and $W_t$ are separated in time, if $k$ is large, they may be weakly correlated, making bridge functions more difficult to fit.

**4.3.3. Two Views of Current Observation.** Finally, we consider a different kind of reduction, which is valid when we have two separate views of the observation; that is, we can partition each observation $O_t$ as $O_t = (O_t^{(0)}, O_t^{(1)})$, where $O_t^{(0)} \perp\!\!\!\perp O_t^{(1)} | S_t$. In this case, we can define $U_t = S_t$, $W_t = O_t^{(1)}$, and $Z_t = O_t^{(0)}$. This allows us to evaluate any policy where $\pi_e^{(t)}$ may depend on all of $\tau_t$ except for $O_{0:t}^{(0)}$.

This kind of reduction could be appealing in many settings. First, it may be useful for the same kinds of applications as the previous kind of reduction, as it

**Table 1.** Summary of Different PCI Reductions

| PCI reduction | $Z_t$ | $W_t$ | $U_t$ | $\pi_e^{(t)}$ can take as input |
|---|---|---|---|---|
| Current and previous observations (simple) | $O_{t-1}$ | $O_t$ | $S_t$ | $O_t$ |
| Current and previous observations (extended) | $O_{t-1}, A_{t-1}, R_{t-1}$ | $O_t$ | $S_t$ | $O_t$ |
| Current and $k$-prior observations | $O_{t-k'}$ | $O_t$ | $S_t, S_{t-k'+1}$ | $(O_t, \tau_{t-1}) \setminus \tau_{t-k'}$ |
| Two views of current observations | $O_t^{(0)}$ | $O_t^{(1)}$ | $S_t$ | $(O_t, \tau_{t-1}) \setminus O_{0:t}^{(0)}$ |

*Notes.* For each, we provide the explicit reduction in terms of the triplet $(Z_t, W_t, U_t)$, and we summarize what kinds of policies can be evaluated under the respective reduction. For the third row, recall that $k' = \min(k, t)$, and for the fourth row, recall that $O_t = (O_t^{(0)}, O_t^{(1)})$, where $O_t^{(0)} \perp\!\!\!\perp O_t^{(1)} | S_t$.

allows us to consider policies defined on a history of past observations without incurring the costs of the same costs in terms of satisfying Assumption 2 or estimating bridge functions. This reduction could be particularly useful when there are some observation variables that cannot be used directly for decision making. For example, in the personalized education example considered in Section 1, there may be certain testing-based metrics that were specifically collected with the logged data, but that would not be available when a policy was deployed. Similarly, in robotics settings as discussed earlier, there may be cheap sensors that are always available and expensive sensors that are only available in the logged data (Pan et al. 2020). In this case, we could include all such unavailable covariates in $O_t^{(0)}$, and the remaining covariates in $O_t^{(1)}$, and this would allow policy evaluation with no effective restriction on the kinds of policies considered. Similarly, if certain sensitive covariates were not allowed to be included in policies, for example, for ethical reasons, such covariates could be included in $O_t^{(0)}$.

### 4.4. Example: Tabular POMDPs Using Previous and Current Observation as Proxies

Finally, we conclude this section with a discussion of our key identification assumptions for a simple tabular case, where we use the previous and current observations as proxies for the unobserved state as described in Section 4.3.1. That is, we consider settings where $U_t = S_t$, $Z_t = O_{t-1}$, $W_t = O_t$, and $\mathcal{S}$ and $\mathcal{O}$ are both finite.

As argued previously, this choice of proxies satisfies Assumption 1 as long as $\pi_e^{(t)}$ depends on $O_t, \tau_{t-1}$ via $O_t$ only. However, it remains to also justify Assumption 2. The following proposition allows us to rewrite the bridge equations for this simple setting in terms of some conditional probability matrices under the POMDP and evaluation policy $\pi_e$.

**Proposition 1.** *Let $P^{(t)}(\mathbf{O}|\mathbf{S})$ denote the $|\mathcal{O}|$ by $|\mathcal{S}|$ matrix of the distribution of $O_t$ given $S_t$ in the POMDP, and let $P_e^{(t)}(\mathbf{S}'|\mathbf{S})$ denote the $|\mathcal{S}|$ by $|\mathcal{S}|$ matrix of the distribution of $S_{t-1}$ given $S_t$ under rollout by $\pi_e$. In addition, for any outcome variable $Y_t = \phi(R_t, D_{t+1:H})$ and $a \in \mathcal{A}$, let $\mathbb{E}_t^*[\mathbb{1}\{E_t = A_t\} Y_t | \mathbf{S}, a]$ denote the $|\mathcal{S}|$-length vector of values of $\mathbb{1}\{E_t = A_t\} Y_t$ given $S_t$ and $A_t = a$ under $\mathcal{P}_t^*$, and let $P_t^*(a|\mathbf{S})^{-1}$ denote the $|\mathcal{S}|$-length vector of values of $P(A_t = a|S_t)^{-1}$ under $\mathcal{P}_t^*$. Then, using proxies $Z_t = O_{t-1}$ and $W_t = O_t$, and confounders $U_t = S_t$, the bridge equations in Assumption 2 for each t correspond to solving*

$$P_e^{(t)}(\mathbf{S}'|\mathbf{S})^\top P^{(t)}(\mathbf{O}|\mathbf{S})^\top q^{(t)}(\mathbf{O}, a) = P_t^*(a|\mathbf{S})^{-1} \qquad \forall a \in \mathcal{A}$$

*and*

$$P^{(t)}(\mathbf{O}|\mathbf{S})^\top h^{(t,\phi)}(\mathbf{O}, a) = \mathbb{E}_t^*[\mathbb{1}\{E_t = A_t\} Y_t | \mathbf{S}, a] \qquad \forall a \in \mathcal{A},$$

*where $q^{(t)}(\mathbf{O}, a)$ and $h^{(t,\phi)}(\mathbf{O}, a)$ are the $|\mathcal{O}|$-length vector of values of $q^{(t)}(Z_t, a)$ and $h^{(t,\phi)}(W_t, a)$, respectively.*

This proposition follows trivially by applying the fact that $Z_t = O_{t-1}$, $W_t = O_t$, and $U_t = S_t$, and explicitly expanding out the conditional expectations in the bridge equations in terms of $P_e^{(t)}(\mathbf{S}'|\mathbf{S})$ and $P^{(t)}(\mathbf{O}|\mathbf{S})$ given the Markovian property of the POMDP conditioned on the unobserved states.

A trivial corollary of the proposition is that, if $|\mathcal{O}| \geq |\mathcal{S}|$, and $P^{(t)}(\mathbf{O}|\mathbf{S})$ and $P_e^{(t)}(\mathbf{S}'|\mathbf{S})$ are both full rank, then the previous equations are always solvable for all $a \in \mathcal{A}$, no matter the outcome variable $Y_t$. This follows by using any pseudo-inverse for $P_e^{(t)}(\mathbf{S}'|\mathbf{S})^\top$ $P^{(t)}(\mathbf{O}|\mathbf{S})^\top$ and $P^{(t)}(\mathbf{O}|\mathbf{S})^\top$. The conditions that $|\mathcal{O}| \geq |\mathcal{S}|$ and that $P^{(t)}(\mathbf{O}|\mathbf{S})$ is full rank are independent of the behavior or evaluation policies, and they essentially require that all distributions over states imply different distributions over observations; that is, there are no "invisible" aspects of $S_t$ that don't affect $O_t$. Conversely, the assumption that $P_e^{(t)}(\mathcal{S}'|\mathcal{S})$ is full rank depends on the evaluation policy $\pi_e$. However, it may be justified for *all* possible evaluation policies, for example, if the $|\mathbf{S}|$ by $|\mathbf{S}|$ conditional probability matrix defining the transition kernel $P_T^{(t)}(S_t | S_{t-1}, A_{t-1} = a)$ were invertible for every $a \in \mathcal{A}$. In other words, we can justify Assumption 2 under some basic conditions on the underlying POMDP, which may be reasoned about on a problem-by-problem basis.

Finally, although the previous analysis is specific to our example setting, the intuition is very general; for Assumption 2 to hold, we need that the proxies are sufficiently well correlated with the confounders (e.g., that $P^{(t)}(\mathbf{O}|\mathbf{S})$ and $P_e^{(t)}(\mathbf{S}'|\mathbf{S})$ are full rank), and that they contain at least as much information as the confounders (e.g., that we also have $|\mathcal{O}| \geq |\mathcal{S}|$).

## 5. Policy Value Estimators

Now we turn from the question of identification to that of estimation. We will focus on estimation of $v_\gamma(\pi_e)$ based on the identification result given by Corollary EC.1 in the online appendix. We will assume in the remainder of this section that we have fixed a valid PCI reduction that satisfies Assumptions 1 and 2. A natural approach to estimating $v_\gamma(\pi_e)$ based on Corollary EC.1 in the online appendix would be to use an estimator of the kind

$$\hat{v}_\gamma^{(n)}(\pi_e) = \frac{1}{n} \sum_{i=1}^{n} \widehat{\psi}_{\mathrm{DR}}(\tau_H^{(i)}), \tag{6}$$

where $\widehat{\psi}_{\mathrm{DR}}$ is an approximation of $\psi dr$ using plug-in estimators for the nuisance functions $h^{(t)}$ and $q^{(t)}$ for each $t$. Specifically, to eschew assumptions on the nuisance function estimators aside from rates, we will use a cross-fitting estimation technique (Zheng and van der Laan 2011, Chernozhukov et al. 2018). Namely, fixing $K \geq 2$, for each $k = 1, \ldots, K$: (1) for $t = 1, \ldots, H$, we fit estimators $\hat{h}^{(t,k)}$ and

$\hat{q}^{(t,k)}$ only on the observed trajectories $i = 1, \dots, n$ with $i \neq k - 1 \,(\mathrm{mod}\, K)$; (2) and then for $i = 1, \dots, n$ with $i = k - 1 \,(\mathrm{mod}\, K)$, we set $\hat{\psi}_{\mathrm{DR}}(\tau_H^{(i)})$ to be $\psi_{\mathrm{DR}}(\tau_H^{(i)})$ where we replace $h^{(t)}, q^{(t)}$ with $\hat{h}^{(t,k)}, \hat{q}^{(t,k)}$. Then we use these to construct an estimator by taking an average as in Equation (6). We discuss exactly how we fit nuisance estimators given trajectory data in Section 5.3. Until then, for Sections 5.1 and 5.2, we keep this abstract and general: We will only impose assumptions about the rates of convergence of nuisance estimators and that we used cross-fitting so that $\tau_H^{(i)}$ is independent of $\hat{h}^{(t,k)}, \hat{q}^{(t,k)}$ whenever $i = k - 1 \,(\mathrm{mod}\, K)$.

### 5.1. Consistency and Asymptotic Normality

We first consider conditions under which the estimator $\hat{v}_\gamma^{(n)}(\pi_e)$ is consistent and asymptotically normal. For this, we need to make some assumptions on the quality of our nuisance estimators.

**Assumption 3.** *Consistent and bounded nuisance estimates*: *letting* $\Psi$ *represent any of* $\{q^{(t)}(Z_t, A_t) : t \in [H]\} \cup \{h^{(t)}(W_t, a) : t \in [H], a \in \mathcal{A}\}$, *we have that for each* $k \in [K]$, (1) $\|\hat{\Psi}^{(k)} - \Psi\|_{2,\mathcal{P}_b} = o_p(1)$, (2) $\|\hat{\Psi}^{(k)}\|_\infty = O_p(1)$, *and* (3) $\|\Psi\|_\infty < \infty$.

*Nuisance estimation rates:* (1) *for each* $t \in [H], a \in \mathcal{A}$, $k \in [K]$, $\|\hat{q}^{(t,k)}(Z_t, A_t) - q^{(t)}(Z_t, A_t)\|_{2,\mathcal{P}_b} \|\hat{h}^{(t,k)}(W_t, a) - h^{(t)}(W_t, a)\|_{2,\mathcal{P}_b} = o_p(n^{-1/2})$, (2) *for each* $t \in [H], t' < t, a \in \mathcal{A}$, $k \in [K]$, $\|\hat{q}^{(t',k)}(Z_{t'}, A_{t'}) - q^{(t')}(Z_{t'}, A_{t'})\|_{2,\mathcal{P}_b} \|\hat{h}^{(t,k)}(W_t, a) - h^{(t)}(W_t, a)\|_{2,\mathcal{P}_b} = o_p(n^{-1/2})$, *and* (3) *for each* $t \in [H], t' < t$, $k \in [K]$, $\|\hat{q}^{(t',k)}(Z_{t'}, A_{t'}) - q^{(t')}(Z_{t'}, A_{t'})\|_{2,\mathcal{P}_b} \|\hat{q}^{(t,k)}(Z_t, A_t) - q^{(t)}(Z_t, A_t)\|_{2,\mathcal{P}_b} = o_p(n^{-1/2})$. *In all of these, the randomness in each bound is with respect to the sampling distribution of the data.*

Essentially, Assumption 3 requires that the nuisances $q^{(t)}$ and $h^{(t)}$ are estimated consistently in terms of the $L_{2,\mathcal{P}_b}$ functional norm for each $t$ and that the corresponding product-error terms converge faster than $n^{-1/2}$ rate. This could be achieved, for example, if each nuisance by itself were estimated at a $o_p(n^{-1/4})$ rate, which notably permits slower-than-parametric rates and is obtainable for many nonparametric machine-learning-based methods (Chernozhukov et al. 2018). In particular, there is a very established line of work on establishing rates like these for conditional moment problems, like those defining $q^{(t)}$ and $h^{(t)}$, in terms of projected error (e.g., obtaining rates for $\|\mathbb{E}[\hat{h}^{(t,k)}(W_t, A_t) - h^{(t)}(W_t, A_t) | Z_t, A_t]\|_2$) using, for example, sieve methods (Chen and Pouzo 2009, 2012) or minimax methods with general machine learning classes (Dikkala et al. 2020). These can be translated to corresponding rates for the actual $L_2$ error (e.g., $\|\hat{h}^{(t,k)}(W_t, A_t) - h^{(t)}(W_t, A_t)\|_2$) given assumptions on so-called "ill-posedness" measures (Chen and Pouzo 2012), which can be used to ensure our required rates. Alternatively, there exist methods that can directly obtain $L_2$ error rates for such conditional moment problems, by leveraging so-called "source conditions" (Carrasco et al. 2007, definition 3.4), for example, using regularized sieve methods (Florens et al. 2011), neural nets with Tikhonov regularization (Liao et al. 2020) or kernel methods with spectral regularization (Wang et al. 2022). The product-rate condition allows for some trade off where, if some nuisances can be estimated faster, then other nuisances can be estimated even slower than $o_p(n^{-1/4})$. In addition, we require a technical boundedness condition on the uniform norm of the errors and of the true nuisances themselves. Given this, we can now present our main consistency and asymptotic normality theorem.

**Theorem 3.** *Let the conditions of Theorem 2 be given, and assume that the nuisance functions plugged into* $\hat{v}_\gamma^{(n)}(\pi_e)$ *are estimated using cross fitting. Furthermore, suppose that the nuisance estimation for each cross-fitting fold satisfies Assumption 3. Then, we have*

$$\sqrt{n}(\hat{v}_\gamma^{(n)}(\pi_e) - v_\gamma(\pi_e)) \to \mathcal{N}(0, \sigma_{\mathrm{DR}}^2) \quad \textit{in distribution,}$$
$$\textit{where} \quad \sigma_{\mathrm{DR}}^2 = \mathbb{E}_{\mathcal{P}_b}[(\psi_{\mathrm{DR}}(\tau_H) - v_\gamma(\pi_e))^2].$$

The key step in proving Theorem 3 is to establish that $\psi dr$ enjoys Neyman orthogonality with respect to all nuisance functions and in particular characterizing the unique product structure of the bias. Having established this, we proceed by applying the machinery of theorem 3.1 of Chernozhukov et al. (2018). We refer the reader to the appendix for the detailed proof.

One technical note about this theorem is that there may be multiple $q^{(t)}$ and $h^{(t)}$ that solve Equations (2) and (3), which creates some ambiguity in both Assumption 3 and the definition of $\psi_{\mathrm{DR}}(\tau_H)$. This is important because the ambiguity in the definition of $\psi_{\mathrm{DR}}(\tau_H)$ affects the value of the asymptotic variance $\sigma_{\mathrm{DR}}^2$. In this case, we implicitly assume that Assumption 3 holds for some arbitrarily given solutions $q^{(t)}$ and $h^{(t)}$ for each $t \in [H]$, and that $\sigma_{\mathrm{DR}}^2$ is defined using the same $q^{(t)}$ and $h^{(t)}$ solutions. Thus, our consistency result in Theorem 3 holds even when bridge functions are nonunique.

Finally, we briefly consider how this variance grows in terms of $H$. Because $\varphi_{DR}(\tau_H)$ consists of a sum of $H$ terms, each of which is multiplied by $\eta_t = \prod_{s'=0}^{t-1} q^{(s')}(Z_{s'}, A_{s'}) \mathbb{1}\{E_{s'} = A_{s'}\}$, we can generally bound the efficient asymptotic variance by $\sum_{t=1}^{H} \prod_{s=1}^{t} \|q^{(s)}(Z_s, A_s)\|_\infty (\|q^{(t)}(Z_t, A_t)\|_2 + \sum_{a \in \mathcal{A}} \|h^{(t)}(W_t, a)\|_2 + \|q^{(t)}(Z_t, A_t)\|_\infty \|h^{(t)}(W_t, A_t)\|_2)$. Therefore, assuming that all functions $h^{(t)}(W_t, A_t)$ and $q^{(t)}(Z_t, A_t)$ have $\|\cdot\|_\infty$ norm of the same order $H$ grows, the asymptotic variance should grow roughly as $\mathcal{O}(H^2)$ as $H \to \infty$. Conversely, if the inverse problems for $q^{(t)}$ and $h^{(t)}$ grow increasingly ill-conditioned as $t$ increases, then the norms of these functions may grow,

in which case the growth of asymptotic variance may be worse than quadratic.

### 5.2. Semiparametric Efficiency

We now consider the question of *semiparametric efficiency* of our OPE estimators. Semiparametric efficiency is defined relative to a model $\mathcal{M}$, which is a set of allowed distributions such that $\mathcal{P}_b \in \mathcal{M}$. Roughly speaking, we say that an estimator is semiparametrically efficient w.r.t. $\mathcal{M}$ if it is regular (meaning invariant to $O_p(1/\sqrt{n})$ perturbations to the data-generating process that keep it inside $\mathcal{M}$), and achieves the minimum asymptotic variance of all regular estimators. We provide a summary of semiparametric efficiency as it pertains to our results in Section EC.4 in the online appendix, but for the purposes of this section it suffices to say that, under conditions we establish, there exists a function $\psi_{\text{eff}} \in L_{2,\mathcal{P}_b}(\tau_H)$, called the "efficient influence function" w.r.t. $\mathcal{M}$, and that an estimator $\hat{v}_\gamma^{(n)}(\pi_e)$ is efficient w.r.t. $\mathcal{M}$ if and only if $\sqrt{n}(\hat{v}_\gamma^{(n)}(\pi_e) - v_\gamma(\pi_e)) = n^{-1/2}\sum_{i=1}^n \psi_{\text{eff}}(\tau_H^{(i)}) + o_p(1)$, that is, asymptotically it looks like simple sample average of this function.

One complication in considering models of distributions on $\tau_H$ is that technically the definition of $v_\gamma(\pi_e)$ depends on the full distribution of $\tau_H^{\text{full}}$. In the case that the distribution of $\tau_H$ corresponds to the logging distribution induced by some behavior policy and underlying POMDP that satisfies Assumption 2, it is clear from Theorem 2 that using *any* nuisances satisfying the required conditional moments will result in the same policy value estimate $v_\gamma(\pi_e)$. However, if we allow for distributions on $\tau_H$ that do not necessarily satisfy such conditions, as is standard in the literature on policy evaluation, it may be the case that different solutions for $h^{(t)}$ and $q^{(t)}$ result in different values of $\mathbb{E}_{\mathcal{P}}[\psi_{\text{DR}}(\tau_H)]$. To avoid such issues, we consider a model of distributions where the nuisances and corresponding policy value estimate are uniquely defined, as follows.

**Definition 1** (Model and Target Parameter). Define $\mathcal{M}_e^{(0)}$ as the set of all distributions on $\tau_H$, and for each $t \geq 1$ recursively define

1. $\eta_{t,\mathcal{P}} = \prod_{s=1}^{t-1} q_{\mathcal{P}}^{(s)}(Z_s, A_s)\ \mathbb{1}\{A_s = E_s\}$,
2. $P_{t,\mathcal{P}}^*(A_t \mid W_t) = \mathbb{E}_{\mathcal{P}}[\eta_{t,\mathcal{P}} \mid W_t, A_t] P_{\mathcal{P}}(A_t \mid W_t)$,
3. $T_{t,\mathcal{P}} : L_{2,\mathcal{P}}(Z_t, A_t) \to L_{2,\mathcal{P}}(W_t, A_t)$, where $(T_{t,\mathcal{P}} g)(W_t, A_t) = \mathbb{E}_{\mathcal{P}}[\eta_{t,\mathcal{P}} g(Z_t, A_t) \mid W_t, A_t]$,
4. $\mathcal{M}_e^{(t)} = \mathcal{M}_e^{(t-1)} \cap \{\mathcal{P} : T_{t,\mathcal{P}}$ is invertible and $P_{t,\mathcal{P}}^*(A_t \mid W_t)^{-1} \in L_{2,\mathcal{P}}(W_t, A_t)\}$,
5. $q_{\mathcal{P}}^{(t)}(Z_t, A_t) = T_{t,\mathcal{P}}^{-1}(P_{t,\mathcal{P}}^*(A_t \mid W_t)^{-1})$,

where 1–3 are defined for $\mathcal{P} \in \mathcal{M}_e^{(t-1)}$, and 5 for $\mathcal{P} \in \mathcal{M}_e^{(t)}$. Furthermore, let $T_{t,\mathcal{P}}^*$ denote the adjoint of $T_{t,\mathcal{P}}$, define $Y_H = R_h$, and for each $t \in [H]$ and $\mathcal{P} \in \mathcal{M}_e^{(t)}$ recursively define

6. $\mu_{t,\mathcal{P}}(Z_t, A_t) = \mathbb{E}_{\mathcal{P}}[\eta_{t,\mathcal{P}}\ \mathbb{1}\{A_t = E_t\}\ Y_{t,\mathcal{P}} \mid Z_t, A_t]$,
7. $h_{\mathcal{P}}^{(t)}(W_t, A_t) = (T_{t,\mathcal{P}}^*)^{-1}(\mu_{t,\mathcal{P}}(Z_t, A_t))$,
8. $Y_{t-1,\mathcal{P}} = R_{t-1} + \gamma(\sum_{a\in\mathcal{A}} h_{\mathcal{P}}^{(t)}(W_t, a) + q_{\mathcal{P}}^{(t)}(Z_t, A_t)(\mathbb{1}\{A_t = E_t\}\, Y_{t,\mathcal{P}} - h_{\mathcal{P}}^{(t)}(W_t, A_t)))$,

where the latter is only defined for $t > 1$. Finally, let $\mathcal{M}_{\text{PCI}} = \mathcal{M}_e^{(H)}$, and for each $\mathcal{P} \in \mathcal{M}_{\text{PCI}}$ define

$$V(\mathcal{P}) = \mathbb{E}_{\mathcal{P}}\left[\sum_{a\in\mathcal{A}} h_{\mathcal{P}}^{(1)}(W_1, a)\right].$$

This definition is not circular because $\eta_{1,\mathcal{P}} = 1$ for every $\mathcal{P}$, and so we can concretely define the first set of quantities in the order they are listed previously for each $t \in [H]$ in ascending order, and the second set in descending order of $t$. The case that $\mathcal{P} = \mathcal{P}_b$, it is straightforward to reason that $\eta_{t,\mathcal{P}_b}$, $q_{\mathcal{P}_b}^{(t)}$, $h_{\mathcal{P}_b}^{(t)}$, and $Y_{t,\mathcal{P}_b}$ agree with the corresponding definitions in Theorem 2 and Corollary EC.1 in the online appendix $T_{t,\mathcal{P}_b}$ and $T_{t,\mathcal{P}_b}^*$ correspond to standard conditional expectation operators under $\mathcal{P}_t^*$, $P_{t,\mathcal{P}_b}^*(A_t \mid W_t) = P_t^*(A_t \mid W_t)$, and $V(\mathcal{P}_b) = v_\gamma(\pi_e)$. Therefore, $\mathcal{M}_{\text{PCI}}$ is a natural model of observational distributions where the required nuisances are uniquely defined, and $V(\mathcal{P})$ is a natural and uniquely defined generalization of $v_\gamma(\pi_e)$ for distributions $\mathcal{P}$ that do not necessarily correspond to actual logging distributions satisfying Assumption 2.

Finally, we assume the following the following on the actual observed distribution $\mathcal{P}_b$.

**Assumption 4.** *For every sequence of distributions $\mathcal{P}_n$ that converge in law to $\mathcal{P}_b$, there exists some integer $N$ such that for all $n \geq N$ and $t \in [H]$ such that $T_{t,\mathcal{P}_n}$ and $T_{t,\mathcal{P}_n}^*$ are invertible. Furthermore, for all such sequences and $t \in [H]$, we also have*

1. $\liminf_{n\to\infty} \inf_{\|f(Z_t,A_t)\|_{1,\mathcal{P}_n}\geq 1} \|T_{t,\mathcal{P}_n} f(Z_t, A_t)\|_{1,\mathcal{P}_n} > 0$,
2. $\liminf_{n\to\infty} \inf_{\|g(W_t,A_t)\|_{1,\mathcal{P}_n}\geq 1} \|T_{t,\mathcal{P}_n}^* g(W_t, A_t)\|_{1,\mathcal{P}_n} > 0$,
3. $\limsup_{n\to\infty} \|P_{t,\mathcal{P}_n}^*(A_t \mid W_t)^{-1}\|_\infty < \infty$.

*In addition, for each $t \in [H]$ the distribution $\mathcal{P}_b$ satisfies*

4. $\inf_{\|f(Z_t,A_t)\|_{2,\mathcal{P}_b}\geq 1} \|T_{t,\mathcal{P}_n} f(Z_t, A_t)\|_{2,\mathcal{P}_b} > 0$,
5. $\inf_{\|g(W_t,A_t)\|_{2,\mathcal{P}_n}\geq 1} \|T_{t,\mathcal{P}_n}^* g(W_t, A_t)\|_{2,\mathcal{P}_b} > 0$.

The condition that $T_{t,\mathcal{P}_n}$ and $T_{t,\mathcal{P}_n}^*$ are invertible for large $n$ ensures that the model $\mathcal{M}_{\text{PCI}}$ is locally saturated at $\mathcal{P}_b$, and the additional conditions ensure that the nuisance functions can be uniformly bounded within parametric submodels. These are very technical conditions used in our semiparametric efficiency proof, and it may be possible to relax them. In discrete settings, these conditions follow easily given $\mathcal{P}_b \in \mathcal{M}_{\text{PCI}}$ because in this setting the conditions can be characterized in terms of the entries or eigenvalues of some probability matrices being bounded away from zero, which by continuity must be the case when $\mathcal{P}_n$ is sufficiently close to $\mathcal{P}_b$. Importantly, the locally saturated condition on $\mathcal{M}_{\text{PCI}}$ at $\mathcal{P}_b$ means that the relevant tangent space is unrestricted. (See Section EC.5.1 in the online appendix for a

discussion of issues with the tangent space in past work in the absence of local saturation.)

Given this setup, we can now present our main efficiency result.

**Theorem 4.** *Suppose that $\mathcal{P}_b$ is the observational distribution given by a POMDP and logging policy that satisfies the conditions of Theorem 2, and let Assumption 4 be given. Then, $\psi_{\mathrm{DR}}(\tau_H) - v_\gamma(\pi_e)$ is the efficient influence function for $V(\mathcal{P})$ at $\mathcal{P} = \mathcal{P}_b$.*

Finally, the following corollary combines this result with Theorem 3, which shows that under the same conditions, if the nuisances are appropriately estimated then the resulting estimator will achieve the semiparametric efficiency bound relative to $\mathcal{M}_{\mathrm{PCI}}$.

**Corollary 1.** *Let the conditions of Theorems 3 and 4 be given. Then, the estimator $\hat{v}_\gamma^{(n)}(\pi_e)$ is semiparametrically efficient w.r.t. $\mathcal{M}_{\mathrm{PCI}}$.*

### 5.3. Nuisance Estimation

Finally, we conclude this section with a discussion of how we may actually estimate $q^{(t)}$ and $h^{(t)}$. The conditional moment Equations (2) and (3) defining these nuisances are defined in terms of the intervention distributions $\mathcal{P}_t^*$, which are not directly observable. Therefore, we provide the following lemma, which reframes these as a nested series of conditional moment restrictions under $\mathcal{P}_b$.

**Lemma 1.** *Let the conditions of Theorem 2 be given. Then, for any collection of functions $q^{(1)}, \dots, q^{(H)}$ and $h^{(1)}, \dots, h^{(H)}$, these functions satisfy Equations (2) and (3) for every $t \in [H]$ if and only if for every $t \in [H]$, we have*

$$\mathbb{E}_{\mathcal{P}_b}\left[\eta_t\left(g(W_t, A_t)q^{(t)}(Z_t, A_t) - \sum_{a\in\mathcal{A}} g(W_t, a)\right)\right] = 0 \quad \forall \text{measurable } g$$

*and*

$$\mathbb{E}_{\mathcal{P}_b}[\eta_t f(Z_t, A_t)(h^{(t)}(W_t, A_t) - \mathbb{1}\{E_t = A_t\} Y_t)] = 0 \quad \forall \text{measurable } f,$$

*where $\eta_t$ and $Y_t$ are defined as in Theorem 2.*

We can observe that the moment restrictions defining $q^{(t)}$ for each $t$ depend only on $q^{(t')}$ for $t' < t$, and those defining $h^{(t)}$ for each $t$ depend on $h^{(t')}$ for $t' > t$ and on $q^{(t'')}$ for every $t'' \neq t$. This suggests a natural order for estimating these nuisances, of $q^{(1)}$ through $q^{(H)}$ first, and then $h^{(H)}$ through $h^{(1)}$. We now take this approach, solving an estimate of the continuum of moment conditions in each round. (An alternative approach may be to jointly solve for all $2H$ nuisances together.) Set

$$U^{(q,t)}(q, g) = \hat{\eta}_t\left(g(W_t, A_t)q(Z_t, A_t) - \sum_{a\in\mathcal{A}} g(W_t, a)\right)$$

$$U^{(h,t)}(h, f) = \hat{\eta}_t f(Z_t, A_t)(h(W_t, A_t) - \mathbb{1}\{E_t = A_t\}\hat{Y}_t),$$

where $\hat{\eta}_t$ and $\hat{Y}_t$ are estimated by plugging in the preceding nuisance estimators (in the ordering described previously). Following Bennett and Kallus (2023), the continuum of moment conditions $\{q : \mathbb{E}_{\mathcal{P}_b} U^{(q,t)}(q, g) = 0 \ \ \forall g\}$ or $\{h : \mathbb{E}_{\mathcal{P}_b} U^{(h,t)}(h, f) = 0 \ \forall f\}$ can be efficiently solved using a regularized, variational reformulation of the optimally weighted generalized method of moments (Hansen 1982), known as the variational method of moments (VMM). This gives our following proposed estimators for solving for this nuisance bridge functions.

**Proposition 2.** *Our VMM nuisance estimators for $q^{(1)}, \dots, q^{(H)}$ and $h^{(1)}, \dots, h^{(H)}$ take the form*

$$q^{(t)} = \arg\min_{q\in\mathcal{Q}^{(t)}} \sup_{g\in\mathcal{G}^{(t)}} \mathbb{E}_n[U^{(q,t)}(q, g)] - \frac{1}{4}\mathbb{E}_n[U^{(q,t)}(\tilde{q}_t, g)^2] + \mathcal{R}^{(q,t)}(q) - \mathcal{R}^{(g,t)}(g),$$

$$h^{(t)} = \arg\min_{h\in\mathcal{H}^{(t)}} \sup_{f\in\mathcal{F}^{(t)}} \mathbb{E}_n[U^{(h,t)}(h, f)] - \frac{1}{4}\mathbb{E}_n[U^{(h,t)}(\tilde{h}_t, f)^2] + \mathcal{R}^{(h,t)}(h) - \mathcal{R}^{(f,t)}(f),$$

*and can be sequentially solved for in the order $q^{(1)}$ through $q^{(H)}$ then $h^{(H)}$ through $h^{(1)}$, where $\mathcal{Q}^{(t)}$ and $\mathcal{H}^{(t)}$ are hypothesis classes for the functions $q^{(t)}$ and $h^{(t)}$, respectively, $\mathcal{G}^{(t)}$ and $\mathcal{F}^{(t)}$ are some critic function classes corresponding to the set of moments we are enforcing, $\mathcal{R}^{(q,t)}$, $\mathcal{R}^{(g,t)}$, $\mathcal{R}^{(h,t)}$, and $\mathcal{R}^{(f,t)}$ are regularizers, and $\tilde{q}^{(t)}$ and $\tilde{h}^{(t)}$ are some prior estimates of $q^{(t)}$ and $h^{(t)}$, which are arbitrarily defined and need not necessarily be consistent.*

There are many existing methods for solving empirical minimax equations of these kinds for different kinds of function classes $\mathcal{Q}^{(t)}$ and $\mathcal{H}^{(t)}$, as well as different kinds of corresponding critic classes $\mathcal{G}^{(t)}$ and $\mathcal{F}^{(t)}$. For example, see Bennett and Kallus (2023) for a detailed description of how such estimators may be implemented for both kernel and neural classes. In particular, in Section EC.6 of the online appendix, we provide a detailed derivation and description of an efficient process for solving these equations when the two critic classes are given by reproducing kernel Hilbert spaces (RKHSs), and we regularize them using squared RKHS norm, which is very generic and allows for any function classes $\mathcal{Q}^{(t)}$ and $\mathcal{H}^{(t)}$ that we can efficiently minimize convex losses over.

## 6. Experiments

Finally, we present a series of experiments to demonstrate our method and theory. We present two sets of experiments. First, we present a simple toy scenario, where we explore the behavior of the methodology and provide a "proof of concept" of our theory. Second, motivated by the findings of our first experiments, we benchmark our methodology in a confounded variation of the more complex "sepsis simulator" environment of Oberst and Sontag (2019), which is a better reflection of

real application. For full details of all experiments, see our code at https://github.com/CausalML/ProximalRL.

### 6.1. Experiment 1: Toy Scenario

**6.1.1. Experimental Setup.** For our first experiment, we consider a simple POMDP, which we refer to as NoisyObs, which is a time-homogeneous POMDP with three states, two actions, and three observation values. We denote these by $\mathcal{S}=\{s_1,s_2,s_3\}$, $\mathcal{A}=\{a_1,a_2\}$, and $\mathcal{O}=\{o_1,o_2,o_3\}$. We detail the state transition, reward, and initial state distribution of the POMDP in Section EC.7 of the online appendix. The observation emission process for NoisyObs is given $P_O^{(t)}(o_i|s_j)=\mathbb{1}\{i=j\}(1-3\epsilon_{\text{noise}}/2)+\epsilon/2$, where $\epsilon_{\text{noise}}$ is a parameter of the POMDP. This models a noisy observation of the state, because we observe the correct state with probability $1-\epsilon_{\text{noise}}$, or a randomly selected incorrect state otherwise. Thus, if $\epsilon_{\text{noise}}=0$ there is no confounding, and greater $\epsilon_{\text{noise}}$ indicate more noisy measurements.

We collected logged data using a time-homogeneous behavioral policy $\pi_b^{\text{NoisyObs}}$, with a horizon length $H=3$. We considered three different evaluation policies $\pi_e^{\text{easy}}$, $\pi_e^{\text{hard}}$, and $\pi_e^{\text{optim}}$, which are all also time-homogeneous and depend only on the current observation and are detailed in Section EC.7 of the online appendix. These polices are so named because $\pi_e^{\text{easy}}$ and $\pi_e^{\text{hard}}$ are designed to have high and low overlap with the logging policy, respectively, and $\pi_e^{\text{optim}}$ is the optimal policy when $\epsilon_{\text{noise}}$ is sufficiently small. Therefore, these cover a wide range of different kinds of policies. In all cases, we set $\gamma=1$.

We performed policy evaluation with the following methods: (1) Ours is the efficient estimator discussed in Section 5, with nuisance estimation performed using the sequential procedure described in Section 5.3; (2) MeanR is a naive unadjusted baseline given by $\frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\gamma^t R_t^{(i)}$; (3) MDP is a model-based baseline given by fitting a tabular MDP to the observed data, treating the observations as states, and computing the value of $\pi_e$ on this model; and (4) TIS is a baseline based on the result in Theorem 1, with estimated plugged-in nuisances and replacing the expectation under $\mathcal{P}_{\text{ind}}$ with its empirical analogue. We provide more detail about each of these methods in Section EC.7 of the online appendix. In the case of our method, we used a simplified version of the "current and previous observation" PCI reduction given by the first row of Table 1, where $Z_t=O_{t-1}$ and $W_t=O_t$, which is valid because we are considering evaluation policies that only depend on $O_t$.

**6.1.2. Results.** We now present results policy evaluation for the previous scenario and policies, using both our method and the previous benchmarks. Specifically, for each $n\in\{200,500,1,000,2,000,5,000,10,000\}$, $\pi_e\in\{\pi_e^{\text{easy}},\pi_e^{\text{hard}},\pi_e^{\text{optim}}\}$, and $\epsilon\in\{0,0.2\}$, we repeated the following process 100 times: (1) we sampled $n$ trajectories with horizon length $H=3$, behavior policy $\pi_b^{\text{NoisyObs}}$ and noise level $\epsilon_{\text{noise}}=\epsilon$; and (2) estimated $v_1(\pi_e)$ using these $n$ trajectories for each method.

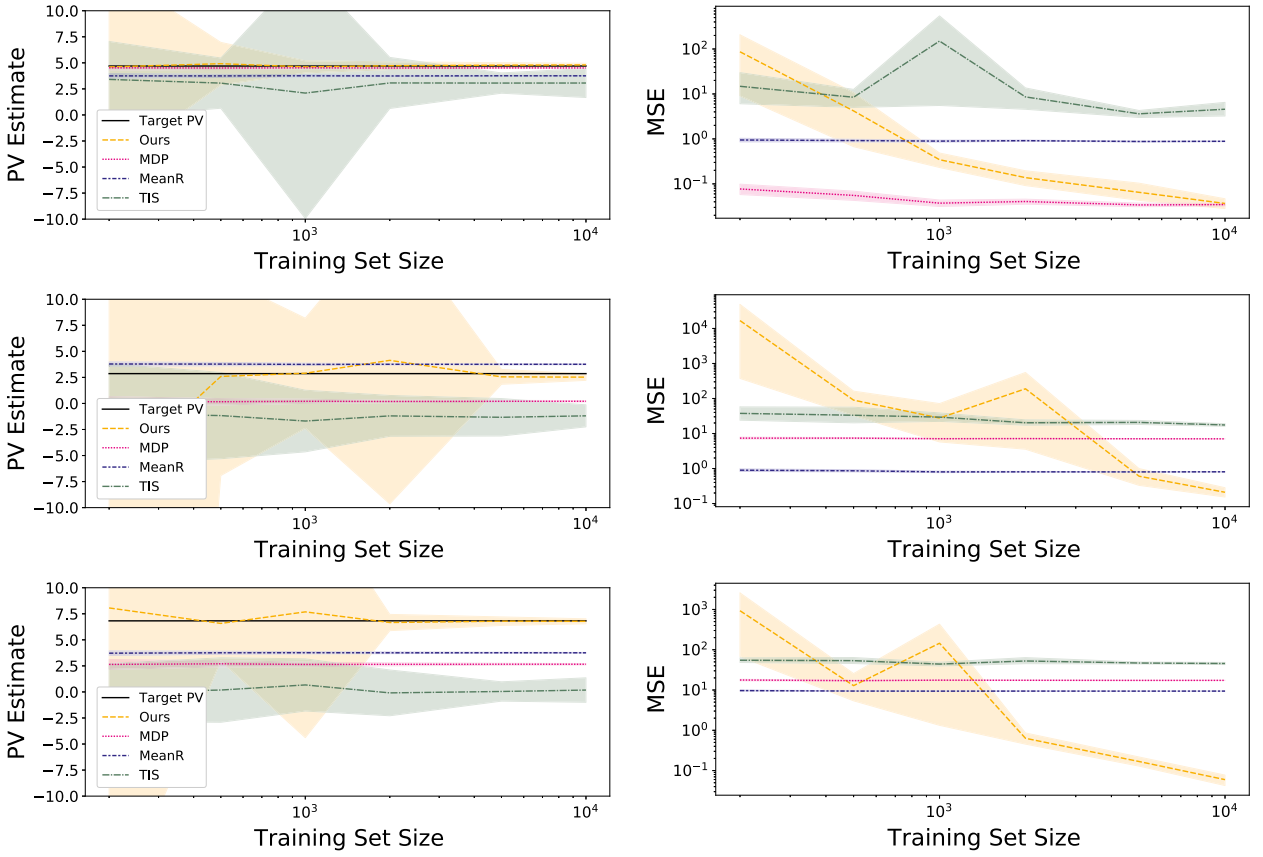
In Figure 4, we display results for the confounded case where $\epsilon_{\text{noise}}=0.2$ (i.e., POMDP setting). Here, we see that our method is consistent, whereas the MDP method, which is only designed to work in MDP settings, is not. The only exception is for estimating the value of $\pi_e^{\text{easy}}$; however, this is only because MDP just happens to have very small bias for estimating this policy. Although our method is consistent, it does have more variance than the MDP benchmark as it tackles a much more complex estimation problem. As expected, the unadjusted MeanR benchmark is inconsistent as it only estimates the value of the logging policy. Finally, despite our identification theory in Section 4.1, the TIS method in general performs very poorly. This is unsurprising, because as discussed in Section 4.1, the identification result (as an expectation over $\mathcal{P}_{\text{ind}}$) may not lend itself to good estimation by plugging in empirical estimates into the identification formula. For comparison, in Section EC.7 of the online appendix, we present additional results for the *un*confounded case, $\epsilon_{\text{noise}}=0$ (i.e., MDP setting), where we see that the MDP baseline becomes consistent due to the absence of confounding and that our method remains consistent and has less variance than in the POMDP setting shown here but still more than the MDP baseline, which is expected as it still solves a more complex estimation problem to adapt to both the MDP and POMDP settings.

### 6.2. Experiment 2: Sepsis Management

**6.2.1. Experimental Setup.** Next, we consider a more "real world"-inspired scenario. Specifically, we consider a scenario based on the sepsis management simulator of Oberst and Sontag (2019). Their environment considers the active management of sepsis for patients, whose state is described by heart rate, blood pressure, oxygen concentration, glucose level, and whether the patient is diabetic. At each time step, the action taken consists of three binary components: whether to place the patient on/off antibiotics, whether to place them on/off vasopressors, and whether to place them on/off a ventilator, giving a total of eight unique actions. After taking each action, we receive a reward based on the number of components of the state taking values within safe ranges, with a maximum reward of 1 if all indicators are safe and the patient is off all three treatments and a minimum reward of $-10$ if three more indicators are unsafe, with various intermediate values. The system uses almost identical parameters as in Oberst and Sontag (2019) with some minor modifications, and we provide a more detailed description in the online appendix.

To introduce confounding, we only observe a censored version of the state; for each patient, with 25% probability, we do not observe whether that patient is diabetic (i.e., in all observations for that patient the

**Figure 4.** (Color online) Experiment Results with $\epsilon_{\text{noise}} = 0.2$




*Notes.* In the top, middle, and bottom rows we display results for $\pi_e^{\text{easy}}$, $\pi_e^{\text{hard}}$, and $\pi_e^{\text{optim}}$, respectively. On the left, we display the mean policy value estimate of each method, where the solid black line corresponds to the true policy value, and the shaded regions correspond to one standard deviation of the policy value estimates. On the right, we display the corresponding mean squared error of these estimates, where the shaded regions correspond to 95% confidence intervals for these values.

"diabetic" indicator is set to "False" regardless of whether the patient is diabetic or not). That is, the true state contains both an indicator of whether the patient is diabetic or not and whether their diabetes status is censored, but for the observed state we instead only observed a possibly censored diabetes indicator. Because all other components of the state are discrete, this means that both state and observation spaces are discrete (i.e., tabular), with a total state space size of $|\mathcal{S}| = 2,880$, and observation space size of $|\mathcal{O}| = 1,440$.

We experimented on this scenario over a time horizon of $H = 3$ and a discount factor of $\gamma = 1$. We first constructed our behavioral policy $\pi_b$ by computing the optimal policy in the true POMDP $\pi^\star$, and defining $\pi_b$ by introducing $\epsilon$-greedy sampling to $\pi^\star$ with $\epsilon = 0.1$; that is, we defined $\pi_b = 0.9\pi^\star + 0.1\pi_{\text{unif}}$, where $\pi_{\text{unif}}$ is a policy that takes all 8 actions with equal probability. Then, we sampled 10,000 observational trajectories using $\pi_b$, and defined $\pi_e$ to be the predicted optimal policy fit on these trajectories using dynamic programming on a simple count-based tabular MDP model, treating the observations $O_t$ as the true states $S_t$. Because the observations $O_t$ are confounded, we expect that $\pi_e$ should *not* necessarily be an estimate of the actual optimal policy $\pi^\star$.

Next, given the fixed policies $\pi_b$ and $\pi_e$ coming from the first stage of the experiment, we repeated the following procedure 50 times: (1) we sampled 10,000 observational trajectories using $\pi_b$, and (2) we estimated $v_1(\pi_e)$ using those trajectories as input for all methods. We performed policy evaluation with our method, as well as the MeanR and MDP benchmarks, as in the previous experiment. In the case of our method, we experimented with a large range of hyperparameter values, as detailed in the online appendix. In addition, we used the proxies $Z_t = (G_{t-1}, X_t)$ and $W_t = (G_t, X_t)$, where $O_t = (G_t, X_t)$ is a partition of the observation into information about diabetes ($G_t$) and nondiabetes information ($X_t$); see online appendix for more details.

Finally, because we had observed in our prior experiments that our method could be sensitive to hyperparameter values, and also since we lack ground truth so cannot set these "fairly" using, for example, cross-validation, we experimented with the following heuristic procedure automatic hyperparameter selection: (1)

**Table 2.** Results of Our Sepsis Experiments

| Method | $\hat{v}_1(\pi_e)$ | Bias | RMSE | Improvement accuracy |
|---|---|---|---|---|
| Ours (best hyperparameters) | $-2.370 \pm 0.597$ | $-0.096$ | 0.599 | 82% |
| Ours (auto hyperparameters) | $-2.459 \pm 0.182$ | $-0.184$ | 0.258 | 100% |
| MDP | $-1.261 \pm 0.054$ | 1.014 | 1.015 | 0% |
| MeanR | $-1.799 \pm 0.025$ | 0.476 | 0.477 | — |

*Notes.* For reach method, we list the average policy value prediction (with one standard deviation error), along with the empirical bias and root mean squared error. In addition, for each method other than MeanR, we list the method's accuracy of predicting whether $v_1(\pi_e) > v_1(\pi_b)$ or not. For reference, the true policy values were $v_1(\pi_e) = -2.275$ and $v_1(\pi_b) = -1.799$.

we first estimate the policy value using all 81 different possible hyperparameter values; (2) we throw away all estimates that take values outside of the range of observed reward values; and (3) we take the median of the remaining estimates. This heuristic is based on the observation from our prior experiments that, as long as hyperparameter values are within reasonable ranges, our method typically gives estimates that are either fairly accurate, or wildly out-of-bound. We estimated policy value using this heuristic separately for each of the 50 experimental replications.

**6.2.2. Results.** We present the main results of this second experiment in Table 2. There we present results for our method with the single best set of hyperparameters out of all tested (in terms of mean squared error across the 50 replications), as well using the automatic hyperparameter selection heuristic described above. We can first observe that using the single best hyperparameter setup gives policy value predictions that are approximately unbiased, but with very high variance. Qualitatively, this variance seems to be partially explained by unstable predictions in a minority of cases. Conversely, our automatic hyperparameter heuristic gives estimates results in slightly higher bias, but much lower variance, and therefore much lower mean squared error. This strong performance of our heuristic vs. choosing the best single set of hyperparameters is extremely encouraging, since unlike picking a "best" hyperparameter combination, the heuristic is actually feasible in practice, as it does not require any ground truth information for hyperparameter selection. Finally, as in the prior experiments, the benchmark methods, which either do not take into account confounding (MDP), or are completely noncausal (MeanR), both give extremely biased estimates with low variance.

Next, in practice, we are often more concerned about predicting whether $\pi_e$ is an improvement on $\pi_b$ or not rather than the exact policy value of $\pi_e$. Accurately answering this question is important in many applications, where the baseline policy $\pi_b$ reflects current best practices or business as usual, and $\pi_e$ represents a proposed new policy. For example, here we could think of $\pi_b$ representing how physicians currently manage sepsis, and $\pi_e$ as a proposed automated algorithm for sepsis management. We have $v_1(\pi_e) \approx -2.275$ and $v_1(\pi_b) \approx -1.799$, so we would like any method of policy evaluation to be able to correctly predict that the new proposed algorithm ($\pi_e$) is worse than standard physician care ($\pi_b$). Specifically, we evaluate each method by what percentage of the time the policy value estimate is smaller than the observational mean reward (MeanR), as the latter is an unbiased estimate of $v_1(\pi_b)$. We list these results in the final column in Table 2. Our method with the best hyperparameters usually correctly predicts that $\pi_e$ is worse than $\pi_b$, and with our automatic hyperparameter selection heuristic this prediction is *always* correct. On the other hand, the MDP benchmark, which fails to take into account confounding from the censored diabetes measurements, always incorrectly predicts that $\pi_e$ is an improvement on $\pi_b$.

## 7. Conclusion

In this paper, we discussed the problem of OPE in an unknown POMDP as a model for the problem of offline RL with general unobserved confounding. First, we analyzed the recently proposed approach for identifying the policy value for tabular POMDPs (Tennenholtz et al. 2020). We showed that, although it could be placed within a more general framework and extended to continuous settings, it suffers from some theoretical limitations due to the unusual form of the identification formulation, which brings its usefulness for constructing estimators with good theoretical properties into question. Motivated by this, we proposed a new framework for identifying the policy value by sequentially reducing the problem to a series of proximal causal inference problems. Furthermore, we extended this identification framework to a framework of estimators based on double machine learning and cross-fitting (Chernozhukov et al. 2018) and showed that under appropriate conditions such estimators are asymptotically normal and semiparametrically efficient. Finally, we constructed a concrete algorithm for implementing such an estimator and provided an empirical proof of concept of our theory by applying algorithm in a toy synthetic setting with confounding due to noisy measurements, as well as a complex sepsis management setting with confounding due to missing measurements of diabetes.

Perhaps the most significant scope for future work on this topic is in the development of more practical algorithms. Indeed, although our experiments were only intended as a proof of concept of our methods and theory, they also show that our actual proposed estimators can often have high variance even in a simple toy POMDP with a moderate number (e.g., 1,000) of trajectories. There may be ways to improve on this; for example, it may be beneficial to solve the conditional moment problems defining the $q^{(t)}$ and $h^{(t)}$ functions simultaneously rather than sequentially as we proposed, which may result in cascading errors. Another important topic for future work would be to explore hyperparameter optimization strategies, such as the heuristic method we proposed for our sepsis experiments; although we found this heuristic worked well empirically, it may introduce other challenges such as dealing with postselection inference.

Another area where there is significant scope for future work is on the topic of semiparametric efficiency. Extending our model to allow for multiple nuisances, in a way where the parameter of interest is still well defined, is an important open challenge. Additional issues are discussed in Section EC.5.1 of the online appendix.

Finally, in terms of future work, there is the problem of how to actually apply our theory and policy value estimators in real-world sequential decision-making problems involving unmeasured confounding. Although our work is largely theoretical, we hope that it will be impactful in motivating progress toward solving such real-world challenges in practice.

## References

Azizzadenesheli K, Lazaric A, Anandkumar A (2016) Reinforcement learning of POMDPs using spectral methods. *Proc. 29th Conf. on Learn. Theory*, 193–256.

Bennett A, Kallus N (2023) The variational method of moments. *J. Royal Statist. Soc. Ser. B Statist. Methodology* 85(3):810–841.

Bennett A, Kallus N, Li L, Mousavi A (2021) Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. *Proc. Internat. Conf. on Artificial Intelligence and Statist.* (PMLR, New York), 1999–2007.

Bhattacharya S, Badyal S, Wheeler T, Gil S, Bertsekas D (2020) Reinforcement learning for POMDP: Partitioned rollout and policy iteration with application to autonomous sequential repair problems. *IEEE Robotic Automated Lett.* 5(3):3967–3974.

Carrasco M, Florens J-P, Renault E (2007) Linear inverse problems in structural econometrics estimation based on spectral decomposition and regularization. Heckman J, Leamer E, eds. *Handbook of Econometrics*, vol. 6 (Elsevier, Amsterdam), 5633–5751.

Chandak Y, Niekum S, da Silva B, Learned-Miller E, Brunskill E, Thomas PS (2021) Universal off-policy evaluation. *Adv. Neural Inform. Processing Systems* 34:27475–27490.

Chen S, Zhang B (2023) Estimating and improving dynamic treatment regimes with a time-varying instrumental variable. *J. Royal Statistical Society Series B: Statistical Methodology* 85(2):427–453.

Chen X, Pouzo D (2009) Efficient estimation of semiparametric conditional moment models with possibly nonsmooth residuals. *J. Econometrics* 152(1):46–60.

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

Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/debiased machine learning for treatment and structural parameters. *Econometrics J.* 21(1):C1–C68.

Cui Y, Pu H, Shi X, Miao W, Tchetgen Tchetgen EJ (2020) Semiparametric proximal causal inference. Preprint, submitted November 17, https://arxiv.org/abs/2011.08411.

Dikkala N, Lewis G, Mackey L, Syrgkanis V (2020) Minimax estimation of conditional moment models. *Adv. Neural Inform. Processing Systems* 33:12248–12262.

Florens J-P, Johannes J, Van Bellegem S (2011) Identification and estimation by penalization in nonparametric instrumental regression. *Econometric Theory* 27(3):472–496.

Gasse M, Grasset D, Gaudron G, Oudeyer P-Y (2021) Causal reinforcement learning using observational and interventional data. Preprint, submitted June 28, https://arxiv.org/abs/2106.14421.

Ghassami A, Ying A, Shpitser I, Tchetgen ET (2022) Minimax kernel machine learning for a class of doubly robust functionals with application to proximal causal inference. *Proc. Internat. Conf. on Artificial Intelligence and Statist.* (PMLR, New York), 7210–7239.

Hansen LP (1982) Large sample properties of generalized method of moments estimators. *Econometrica* 50(4):1029–1054.

Hu Y, Wager S (2023) Off-policy evaluation in partially observed Markov decision processes under sequential ignorability. Preprint, submitted October 24, https://arxiv.org/abs/2110.12343.

Kallus N, Uehara M (2020) Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *J. Machine Learn. Res.* 21(1):6742–6804.

Kallus N, Uehara M (2022) Efficiently breaking the curse of horizon in off-policy evaluation with double reinforcement learning. *Oper. Res.* 70(6):3035–3628.

Kallus N, Zhou A (2020) Confounding-robust policy evaluation in infinite-horizon reinforcement learning. *Adv. Neural Inform. Processing Systems* 33:22293–22304.

Kallus N, Mao X, Uehara M (2022) Causal inference under unmeasured confounding with negative controls: A minimax learning approach. Preprint, submitted March 25, https://arxiv.org/abs/2103.14029.

Katt S, Oliehoek FA, Amato C (2017) Learning in POMDPs with Monte Carlo tree search. *Proc. Internat. Conf. on Machine Learn.* (PMLR, New York), 1819–1827.

Killian TW, Ghassemi M, Joshi S (2022) Counterfactually guided policy transfer in clinical settings. *Proc. Conf. on Health, Inference, and Learn.* (PMLR, New York), 5–31.

Liao L, Fu Z, Yang Z, Kolar M, Wang Z (2021) Instrumental variable value iteration for causal offline reinforcement learning. Preprint, submitted February 19, https://arxiv.org/abs/2102.09907.

Liao L, Chen Y-L, Yang Z, Dai B, Kolar M, Wang Z (2020) Provably efficient neural estimation of structural equation models: An adversarial approach. *Adv. Neural Inform. Processing Systems* 33: 8947–8958.

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

Miao W, Shi X, Tchetgen Tchetgen EJ (2018b) A confounding bridge approach for double negative control inference on causal effects. Preprint, submitted August 15, https://arxiv.org/abs/1808.04945.

Nair Y, Jiang N (2021) A spectral approach to off-policy evaluation for POMDPs. Preprint, submitted September 22, https://arxiv.org/abs/2109.10502.

Namkoong H, Keramati R, Yadlowsky S, Brunskill E (2020) Off-policy policy evaluation for sequential decisions under unobserved confounding. *Adv. Neural Inform. Processing Systems* 33:18819–18831.

Oberst M, Sontag D (2019) Counterfactual off-policy evaluation with Gumbel-max structural causal models. *Proc. Internat. Conf. on Machine Learn.* (PMLR, New York), 4881–4890.

Pan Y, Cheng C-A, Saigol K, Lee K, Yan X, Theodorou EA, Boots B (2020) Imitation learning for agile autonomous driving. *Internat. J. Robotics Res.* 39(2–3):286–302.

Shi X, Miao W, Nelson JC, Tchetgen Tchetgen EJ (2020) Multiply robust causal inference with double-negative control adjustment for categorical unmeasured confounding. *J. Royal Statist. Soc. Ser. B Statist. Methodology* 82(2):521–540.

Singh G, Peri S, Kim J, Kim H, Ahn S (2021) Structured world belief for reinforcement learning in POMDP. *Proc. 38th Internat. Conf. on Machine Learn.* (PMLR, New York), 9744–9755.

Tchetgen Tchetgen EJ, Ying A, Cui Y, Shi X, Miao W (2020) An introduction to proximal causal learning. Preprint, submitted September 23, https://arxiv.org/abs/2009.10982.

Tennenholtz G, Mannor S, Shalit U (2020) Off-policy evaluation in partially observable environments. *Proc. 33rd AAAI Conf. on Artificial Intelligence* (Curran Associates, Red Hook, NY), 10276–10283.

Wang L, Yang Z, Wang Z (2021) Provably efficient causal reinforcement learning with confounded observational data. *Adv. Neural Inform. Processing Systems* 34:21164–21175.

Wang Z, Luo Y, Li Y, Zhu J, Schölkopf B (2022) Spectral representation learning for conditional moment models. Preprint, submitted October 29, https://arxiv.org/abs/2210.16525.

Xu L, Kanagawa H, Gretton A (2021) Deep proxy causal learning and its application to confounded bandit policy evaluation. *Adv. Neural Inform. Processing Systems* 34:26264–26275.

Yang C-HH, Hung I, Danny T, Ouyang Y, Chen P-Y (2021) Causal inference q-network: Toward resilient reinforcement learning. Preprint, submitted February 18, https://arxiv.org/abs/2102.09677.

Ying A, Miao W, Shi X, Tchetgen Tchetgen EJ (2021) Proximal causal inference for complex longitudinal studies. Preprint, submitted September 15, https://arxiv.org/abs/2109.07030.

Zheng W, van der Laan MJ (2011) Cross-validated targeted minimum-loss-based estimation. *Targeted Learning* (Springer, Berlin), 459–474.

**Andrew Bennett** is a recent PhD graduate from the Department of Computer Science at Cornell University. He is interested in various problems at the intersection of machine learning, causal inference, and econometrics, especially in challenging settings involving unobserved confounding.

**Nathan Kallus** is an associate professor in the School of Operations Research and Information Engineering and Cornell Tech at Cornell University. His research interests include optimization, especially under uncertainty and informed by data; causal inference; sequential decision making; and algorithmic fairness.