# Off-Policy Evaluation in Partially Observable Environments

**Guy Tennenholtz**
Technion Institute of Technology

**Shie Mannor**
Technion Institute of Technology

**Uri Shalit**
Technion Institute of Technology

**Abstract**

This work studies the problem of batch off-policy evaluation for Reinforcement Learning in partially observable environments. Off-policy evaluation under partial observability is inherently prone to bias, with risk of arbitrarily large errors. We define the problem of off-policy evaluation for Partially Observable Markov Decision Processes (POMDPs) and establish what we believe is the first off-policy evaluation result for POMDPs. In addition, we formulate a model in which observed and unobserved variables are decoupled into two dynamic processes, called a Decoupled POMDP. We show how off-policy evaluation can be performed under this new model, mitigating estimation errors inherent to general POMDPs. We demonstrate the pitfalls of off-policy evaluation in POMDPs using a well-known off-policy method, Importance Sampling, and compare it with our result on synthetic medical data.

## 1 Introduction

Reinforcement Learning (RL) algorithms learn to maximize rewards by analyzing past experience in an unknown environment (Sutton and Barto 1998). In the context of RL, off-policy evaluation (OPE) refers to the task of estimating the value of an *evaluation policy* without applying it, using data collected under a different *behavior policy* (Dann, Neumann, and Peters 2014), also unknown as a logging policy.

The problem of OPE has been thoroughly studied under fully-observable models. In this paper we extend and define OPE for Partially Observable Markov Decision Processes (POMDPs). Informally, the goal of OPE in POMDPs is to evaluate the cumulative reward of an *evaluation* policy $\pi_e$ which is a function of *observed* histories, using a measure over the observed variables under a *behavior* policy $\pi_b$ which is a function of an *unobserved* state. We assume that we do not have access to the unobserved states, nor do we have any prior information of their model. In fact, in many cases we do not even know whether these states exist. These states are commonly referred to as confounding variables in the causal inference literature, whenever they affect both the reward as well as the behavior policy. OPE for POMDPs is highly relevant to real-world applications in fields such as healthcare, where we are trying to learn from observed policies enacted by medical experts, without having full access to the information the experts have in hand.



A basic observation we make is that **traditional methods in OPE are not applicable to partially observable environments**. For this reason, we start by defining the OPE problem for POMDPs, proposing various OPE approaches. We define OPE for POMDPs in Section 2. Then, in Section 3, Theorem 1 shows how past and future observations of an unobserved state can be leveraged in order to evaluate a policy, under a non-singularity condition of certain joint probability distribution matrices. To the best of our knowledge, this is the first OPE result for POMDPs.

In Section 4 we build upon the results of Section 3 and propose a more involved POMDP model: the Decoupled POMDP model. This is a class of POMDPs for which observed and unobserved variables are distinctly partitioned. Decoupled POMDPs hold intrinsic advantages over POMDPs for OPE. The assumptions required for OPE are more flexible than those required in POMDPs, allowing for easier estimation (Theorem 2).

In Section 5 we attempt to answer the question as to why traditional OPE methods fail when parts of the state are unobserved. We emphasize the hardness of OPE in POMDPs through a conventional procedure known as Importance Sampling. We further construct an Importance Sampling variant that can be applied to POMDPs under an assumption about the reward structure. We then compare this variant to the OPE result of Section 4 in a synthetic medical environment (Section 6), showing it is prone to arbitrarily large bias.

Before diving into the subtleties associated with off-policy evaluation in partially observable environments, we provide two examples in which ignoring unobserved variables can lead to erroneous conclusions about the relation between actions and their rewards.

**Medical Treatment:** Consider a physician monitoring a patient, frequently prescribing drugs and applying treatments according to her medical state. Some of the patient's information observed by the physician may be unavailable to us (e.g., the patient's socioeconomic status). A physician might tend to prescribe Drug A for her wealthier patients who can afford it. At the same time, wealthier patients tend to have

better outcomes regardless of the specific drug. As we are unaware of the doctor's inner state for choosing this action, and have no access to the information she is using, a naive model would wrongly deduce that prescribing drug A is more effective than it actually is.

**Autonomous driving:** Consider teaching an autonomous vehicle to drive using video footage of cameras located over intersections. In this scenario, many unobserved variables may be present, including: objects unseen by the camera, the driver's current mood, social cues between drivers and pedestrians, and so on. Naive estimations of policies based on other drivers' behavior may result in catastrophic outcomes. For the purpose of this illustration, let us assume tired drivers tend to be involved in more traffic accidents than non-tired drivers. In addition, suppose non-tired drivers tend to drive faster than tired drivers. We wish to construct a safe autonomous car based on traffic camera footage. Since the tiredness of the driver is unobserved, a naive model might wrongly evaluate a good policy as one that drives fast.

Understanding how to evaluate the effects of actions in the presence of unobserved variables that affect both actions and rewards is the premise of a vast array of work in the field of causal inference. Our present work owes much to ideas presented in the causal inference literature under the name of effect restoration (Kuroki and Pearl 2014). In our work, we build upon a technique introduced by Miao, Geng, and Tchetgen Tchetgen (2018) on causal inference using proxy variables. In the two papers above the unobserved variable is static, and there is only one action taken. Our work deals with dynamic environments with sequential actions and action – hidden-state feedback loops. Surprisingly, while this is in general a harder problem, we show that these dynamics can be leveraged to provide us with multiple noisy views of the unobserved state.

Our work sits at an intersection between the fields of RL and Causal Inference. While we have chosen to use terminology common to RL, this paper could have equivalently been written in causal inference terminology. We believe it is essential to bridge the gap between these two fields, and include an interpretation of our results using causal inference terminology in the appendix.

## 2 Preliminaries

### Partially Observable Markov Decision Process

We consider a finite-horizon discounted Partially Observable Markov Decision Process (POMDP). A POMDP is defined as the 5-tuple $(\mathcal{U}, \mathcal{A}, \mathcal{Z}, \mathcal{P}, O, r, \gamma)$ (Puterman 1994), where $\mathcal{U}$ is a finite state space, $\mathcal{A}$ is a finite action space, $\mathcal{Z}$ is a finite observation space, $P : \mathcal{U} \times \mathcal{U} \times \mathcal{A} \mapsto [0, 1]$ is a transition kernel, $O : \mathcal{U} \times \mathcal{Z} \mapsto [0, 1]$ is the observation function, where $O(u, z)$ denotes the probability $P(z|u)$ of perceiving observation $z$ when arriving in state $u$, $r : \mathcal{U} \times \mathcal{A} \to [0, 1]$ is a reward function, and $\gamma \in (0, 1)$ is the discount factor. A diagram of the causal structure of POMDPs is depicted in Figure 1a.

A POMDP assumes that at any time step the environment is in a state $u \in \mathcal{U}$, an agent takes an action $a \in \mathcal{A}$ and receives a reward $r(u, a)$ from the environment as a result of this action. At any time $t$, the agent will have chosen actions and received rewards for each of the $t$ time steps prior to the current one. The agent's observable history at time $t$ is defined by $h_t^o = (z_0, a_0, \ldots, z_{t-1}, a_{t-1}, z_t)$. We denote the space of observable histories at time $t$ by $\mathcal{H}_t^o$.

We consider trajectories and observable trajectories. A trajectory of length $t$ is defined by the sequence $\tau = (u_0, z_0, a_0, \ldots, u_t, z_t, a_t)$. Similarly, an observable trajectory of length $t$ is defined by the sequence $\tau^o = (z_0, a_0, \ldots, z_{t-1}, a_{t-1}, z_t, a_t)$. We denote the space of trajectories and observable trajectories of length $t$ by $\mathcal{T}_t$ and $\mathcal{T}_t^o$, respectively. Finally, a policy $\pi$ is any stochastic, time-dependent[1] mapping from a measurable set $\mathcal{X}_t \subset \mathcal{T}_t$ to the set of probability measures on the Borel sets of $\mathcal{A}$, denoted by $\mathcal{B}(\mathcal{A})$.

For any time $t$, and trajectory $\tau \in \mathcal{T}_t$, we define the cumulative reward

$$R_t(\tau) = \sum_{k=0}^{t} \gamma^k r(u_k, a_k).$$

The above is also known as the discounted return. Given any policy $\pi$ and initial distribution over states $\mathcal{U}$, denoted by $\nu_0$, we define the expected discounted return at time $L$ by

$$v_L(\pi; \nu_0) = \mathbb{E}\left(R_L(\tau) | u_0 \sim \nu_0, \tau \sim \pi\right).$$

When clear from context, we will assume $\nu_0$ and $L$ are known and fixed and simply write $v(\pi)$.

### Policy Evaluation

Off-policy evaluation (OPE) considers two types of policies: a behavior policy and an evaluation policy, as defined below.

**Definition 1** (behavior and evaluation policies)**.**
*A behavior policy, denoted by $\pi_b^{(t)}$, is a stochastic, time-dependent mapping from states $\mathcal{U}$ to $\mathcal{B}(\mathcal{A})$.* [2]
*An evaluation policy, denoted by $\pi_e^{(t)}$, is a stochastic, time-dependent mapping from observable histories $\mathcal{H}_t^o$ to $\mathcal{B}(\mathcal{A})$.*

For any time step $t$ and policy $\pi$ let $P^\pi(\cdot)$ be the measure over observable trajectories $\mathcal{H}_t^o$, induced by policy $\pi$. We will denote this measure by $P^b, P^e$, whenever $\pi$ is a behavior or evaluation policy, respectively. We are now ready to define off-policy evaluation in POMDPs:

> *The goal of off-policy evaluation in POMDPs is to evaluate $v_L(\pi_e)$ using the measure $P^b(\cdot)$ over observable trajectories $\mathcal{T}_L^o$ and the given policy $\pi_e$.*

This corresponds to the scenario in which data comes from a system which records an agent taking actions based on her own information sources ($u$), and we want to evaluate a policy $\pi_e$ which we learn based only on information available to the learning system ($\tau^o$).

[1] For brevity we sometimes denote policies by $\pi$, though they may depend on time such that $\pi = \pi(t)$.

[2] We consider behavior policies to be functions of unobserved states. This assumption is common in MDPs in which the state is observed, as it is known that there exists a stationary optimal policy that is optimal (Romanovskii 1965). In addition, the unobserved state is assumed to contain all required information for an agent to make an optimal decision.

### Vector Notations

Let $x, y, z$ be random variables accepting values in $\{x_1, \ldots, x_{n_1}\}, \{y_1, \ldots, y_{n_2}\}, \{z_1, \ldots, z_{n_3}\}$, respectively, and let $\mathcal{F}$ be a filtration that includes all information: states, observations, actions, and rewards. We denote by $P(X|y, \mathcal{F})$ the $n_1 \times 1$ column vector with elements $(P(X|y, \mathcal{F}))_i = P(x_i|y, \mathcal{F})$. Similarly we denote by $P(x|Y, \mathcal{F})$ the $1 \times n_2$ row vector with elements $(P(x|Y, \mathcal{F})_i = P(x|y_i, \mathcal{F})$. Note that $P(x|Y, \mathcal{F})P(Y|\mathcal{F}) = P(x|\mathcal{F})$. Finally, let $P(X|Y, \mathcal{F})$ be the $n_1 \times n_2$ matrix with elements $(P(X|Y, \mathcal{F}))_{ij} = P(x_i|y_j, \mathcal{F})$. Note that if $x$ is independent of $z$ given $y$ then we have the matrix equality $P(X|Y, \mathcal{F})P(Y|Z, \mathcal{F}) = P(X|Z, \mathcal{F})$.

We will sometimes only consider subsets of the above matrices. For any index sets $I \subset \{1, \ldots n_1\}, J \subset \{1, \ldots n_2\}$ let $P_{(I,J)}(X|Y, \mathcal{F})$ be the matrix with elements $\left(P_{(I,J)}(X|Y, \mathcal{F})\right)_{ij} = P(x_{I_i}|y_{J_j}, \mathcal{F})$.

## 3 Policy Evaluation for POMDPs

In this section we show how past and future observations can be leveraged in order to create an unbiased evaluation of $\pi_e$ under specific invertibility conditions. It is a generalization of the bandit-type result presented in Miao, Geng, and Tchetgen Tchetgen (2018), where causal effects are inferred in the presence of an unobserved discrete confounder, provided one has two conditionally independent views of the confounder which are non-degenerate (i.e., the conditional probability matrices are invertible). We show how time dynamics readily give us these two conditionally independent views - using the past and future observations as two views of the unobserved state.

For any $\tau^o = (z_0, a_0, \ldots, z_t, a_t) \in \mathcal{T}_t^o$ we define the generalized weight matrices

$$W_i(\tau^o) = P^b(Z_i|a_i, Z_{i-1})^{-1} P^b(Z_i, z_{i-1}|a_{i-1}, Z_{i-2})$$

for $i \geq 1$, and

$$W_0(\tau^o) = P^b(Z_0|a_0, Z_{-1})^{-1} P^b(Z_0).$$

Here, we assume there exists an observation of some time step before initial evaluation (i.e., $t < 0$), which we denote by $z_{-1}$. Alternatively, $z_{-1}$ may be an additional observation that is independent of $z_0$ and $a_0$ given $u_0$. Note that the matrices $W_i$ can be estimated from the observed trajectories of the behavior distribution. We then have the following result.

**Theorem 1** (POMDP Evaluation)**.** *Assume $|\mathcal{Z}| \geq |\mathcal{U}|$ and that $P^b(Z_i|a_i, Z_{i-1})$ are invertible for all $i$ and all $a_i \in \mathcal{A}$. For any $\tau^o \in \mathcal{T}_t^o$ denote*

$$\Pi_e(\tau^o) = \prod_{i=0}^{t} \pi_e^{(i)}(a_i|h_i^o), \qquad \Omega(\tau^o) = \prod_{i=0}^{t} W_{t-i}(\tau^o).$$

*Then*

$$P^e(r_t) = \sum_{\tau^o \in \mathcal{T}_t^o} \Pi_e(\tau^o) P^b(r_t, z_t|a_t, Z_{t-1}) \Omega(\tau^o).$$

*Proof.* See Appendix. □

Having evaluated $P^e(r_t)$ for all $0 \leq t \leq L$ suffices in order to evaluate $v(\pi_e)$. Theorem 1 lets us evaluate $v(\pi_e)$ without access to the unknown states $u_i$. It uses past and future observations $z_t$ and $z_{t-1}$ as proxies of the unknown state. Its main assumptions are that the conditional distribution matrices $P^b(Z_i|a_i, Z_{i-1})$ and $P^b(U_i|a_i, Z_{i-1})$ are invertible. In other words, it is assumed that enough information is transferred from states to observations between time steps. Consider for example the case where $U_i$ and $Z_i$ are both binary, then a sufficient condition for invertibility to hold is to have $p(z_i = 1|z_{i-1} = 1, a_i) \neq p(z_i = 1|z_{i-1} = 0, a_i)$ for all values of $a_i$. A trivial case in which this assumption does not hold is when $\{u_i\}_{0 \leq i \leq L}$ are i.i.d. In such a scenario, the observations $z_i, z_{i-1}$ do not contain enough useful information of the unobserved state $u_i$, and additional independent observations of $u_i$ are needed. In the next section we will show how this assumption can be greatly relaxed under a decoupled POMDP model. Particularly, in the next section we will devise an alternative model for partial observability, and analyze its superiority over POMDPs in the context of OPE.

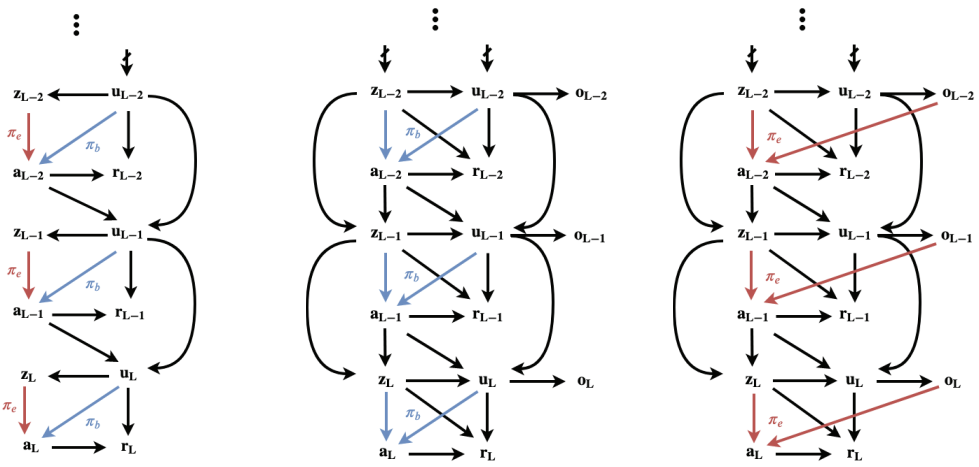
## 4 Decoupled POMDPs

Theorem 1 provides an exact evaluation of $\pi_e$. However it assumes non-singularity of several large stochastic matrices. While such random matrices are likely to be invertible (see e.g., Bordenave, Caputo, and Chafaï (2012) Thm 1.4), estimating their inverses from behavior data can lead to large approximation errors or require very large sample sizes. This is due to the structure of POMDPs, which is confined to a causal structure in which unobserved states form observations. This restriction is present even when $O(u, z)$ is a deterministic measure. In many settings, one may detach the effect of unobserved and observed variables. In this section, we define a Decoupled Partially Observable Markov Decision Process (Decoupled POMDP), in which the state space is decoupled into unobserved and observed variables.

Informally, in Decoupled POMDPs the state space is factored into observed and unobserved states. Both the observed and unobserved states follow Markov transitions. In addition, unobserved states emit independent observations. As we will show, Decoupled POMDPs are more appropriate for OPE under partial observability. Contrary to Theorem 1, Decoupled POMDPs use matrices that scale with the support of the unobserved variables, which, as they are decoupled from observed variables, are of much smaller cardinality.

**Definition 2.** *We define a finite-horizon discounted Decoupled Partially Observable Markov Decision Process (Decoupled POMDP) as the tuple $(\mathcal{U}, \mathcal{Z}, \mathcal{O}, \mathcal{A}, \mathcal{P}, \mathcal{P}_\mathcal{O}, r, \gamma)$, where $\mathcal{Z}$ and $\mathcal{U}$ consist of an observed and unobserved finite state space, respectively. $\mathcal{A}$ is the action space, $\mathcal{O}$ is the independent observation space, $\mathcal{P} : \mathcal{Z} \times \mathcal{U} \times \mathcal{Z} \times \mathcal{U} \times \mathcal{A} \mapsto [0, 1]$ is the transition kernel, where $\mathcal{P}(z', u'|z, u, a)$ is the probability of transitioning to state $(z', u')$ when visiting state $(z, u)$ and taking action $a$, $\mathcal{P}_\mathcal{O} : \mathcal{U} \times \mathcal{O} \mapsto [0, 1]$ is the independent observation function, where $\mathcal{P}_\mathcal{O}(o|u)$ is the probability of receiving observation $o$ when arrive at state $u$, $r : \mathcal{U} \times \mathcal{Z} \times \mathcal{A} \to [0, 1]$ is a reward function, and $\gamma \in (0, 1)$ is the discount factor.*

A Decoupled POMDP assumes that at any time step the environment is in a state $(u, z) \in \mathcal{U} \times \mathcal{Z}$, an agent takes an action $a \in \mathcal{A}$ and receives a reward $r(u, z, a)$ from the environment. The agent's observable history is defined by



(a) POMDP (b) Decoupled POMDP

Figure 1: A causal diagram of a POMDP (a) and a Decoupled POMDP (b). In Decoupled POMDPs, observed and unobserved states are separated into two distinct processes, with a coupling between them at each time step. Diagrams depicts the causal dependence of a behavior policy and evaluation policies. While evaluation policies are depicted to depend on the current observation alone, they can depend on any observable history $h_t^o$.

$h_t^o = (z_0, o_0, a_0, \ldots, z_{t-1}, o_{t-1}, a_{t-1}, z_t, o_t)$. With abuse of notations we denote the space of observable histories at time $t$ by $\mathcal{H}_t^o$. We similarly define trajectories and observable trajectories of length $t$ by $\tau = (u_0, z_0, o_0, a_0, \ldots, u_t, z_t, o_t, a_t)$ and $\tau^o = (z_0, o_0, a_0, \ldots, z_t, o_t, a_t)$, respectively. With abuse of notations, we denote the space of trajectories and observable trajectories of length $t$ by $\mathcal{T}_t$ and $\mathcal{T}_t^o$, respectively. In Figure 1(b) we give an example of a Decoupled POMDP (used in Theorem 2) for which $z_i$ causes $u_i$.

Decoupled POMDPs hold the same expressive power and generality of POMDPs. To see this, one may remove the observed state space $\mathcal{Z}$ to recover the original POMDP model. Nevertheless, as we will see in Theorem 2, Decoupled POMDPs also let us leverage their structure in order to achieve tighter results. Similar to POMDPs, OPE for Decoupled POMDPs considers behavior and evaluation policies.

**Definition 3** (behavior and evaluation policies)**.**
*A behavior policy, denoted by $\pi_b^{(t)}$, is a stochastic, time-dependent mapping from states $\mathcal{U} \times \mathcal{Z}$ to $\mathcal{B}(\mathcal{A})$.*
*An evaluation policy, denoted by $\pi_e^{(t)}$, is a stochastic, time-dependent mapping from observable histories $\mathcal{H}_t^o$ to $\mathcal{B}(\mathcal{A})$.*

The goal of OPE for Decoupled POMDPs is defined similarly as general POMDPs. Decoupled POMDPs allow us to model environments in which observations are not contained in the unknown state. They are decoupled by a Markovian processes which governs both observed and unobserved variables. In what follows, we will show how this property can be leveraged for evaluating a desired policy.

## Policy Evaluation for Decoupled POMDPs

For all $i$, let $K_i \subset \{1, \ldots, |\mathcal{O}|\}, J_i \subset \{1, \ldots, |\mathcal{Z}|\}$ such that $|K_i| = |J_i| = |U|$. Similar to before, for any $\tau^o \in \mathcal{T}_t^o$, we define the generalized weight matrices

$$G_i(\tau^o) = P^b_{(K_i, J_{i-1})}(O_i | z_i, a_i, Z_{i-1})^{-1} \times \\ \times P^b_{(K_i, J_{i-2})}(O_i, o_{i-1}, z_i | z_{i-1}, a_{i-1}, Z_{i-2}),$$

for $i \geq 1$ and

$$G_0(\tau^o) = P^b_{(K_0, J_{-1})}(O_0 | z_0, a_0, Z_{-1})^{-1} P^b_{K_0}(O_0 | z_0) P^b(z_0).$$

We then have the following result.

**Theorem 2** (Decoupled POMDP Evaluation)**.** *Assume $|\mathcal{Z}|, |\mathcal{O}| \geq |\mathcal{U}|$ and that there exist index sets $K_i, J_i$ such that $|K_i| = |J_i| = |\mathcal{U}|$ and $P^b_{(K_i, J_{i-1})}(O_i | a_i, z_i, Z_{i-1})$ are invertible $\forall i, a_i, z_i \in \mathcal{A} \times \mathcal{Z}$. In addition assume that $z_{i-1}$ is independent of $u_{i+1}$ given $z_{i+1}, a_i, u_i$, $\forall i$ under $P^b$.*
*For any $\tau^o \in \mathcal{T}_t^o$, denote by*

$$\Pi_e(\tau^o) = \prod_{i=0}^{t} \pi_e^{(i)}(a_i | h_i^o), \qquad \Omega(\tau^o) = \prod_{i=0}^{t} G_{t-i}(\tau^o).$$

*Then*

$$P^\pi(r_t) = \sum_{\tau^o \in \mathcal{T}_t^o} \Pi_e(\tau^o) P^b_{K_{t-1}}(r_t, o_t | z_t, a_t, Z_{t-1}) \Omega(\tau^o).$$

*Proof.* See Appendix. □

Approximating Theorem 2's result under finite datasets is more robust than Theorem 1. For one, the matrices are of size $|\mathcal{U}|$, which can be much smaller than $|\mathcal{Z}| + |\mathcal{U}|$. In addition, and contrary to Theorem 1, the result holds for any index set $J_i, K_i \subset [|\mathcal{Z}|]$ of cardinality $|\mathcal{U}|$. We can thus choose any of $\binom{|\mathcal{Z}|}{|\mathcal{U}|}$ possible subsets from which to approximate the matrices $G_i(\tau)$. This enables us to choose indices for solutions with desired properties (e.g., small condition numbers of the matrices). We may also construct an estimator based on a majority vote of the separate estimators. Finally, we

note that solving for $G_i(\tau)$ for any $J_i, K_i$ can be done using least squares regression.

Up to this point we have shown how the task of OPE can be carried out in two settings: general POMDPs and Decoupled POMDPs. The results in Theorems 1 and 2 depend on full trajectories, which we believe are a product of the high complexity inherent to the problem of OPE with unobservable states. In the next section, we demonstrate the hardness inherent to OPE in these settings through an alternative OPE method - a variant of an Importance Sampling method (Precup 2000). We then experiment and compare these different OPE techniques on a synthetic medical environment.

## 5 Importance Sampling and its Limitations in Partially Observable Environments

In previous sections we presented OPE in partially observable environments, and provided, for what we believe is the first time, techniques of exact evaluation. A reader familiar with OPE in fully-observable environments might ask, why do new techniques need to be established and where do traditional methods fail? To answer this question, in this section we demonstrate the bias that may arise under the use of long-established OPE methods. More specifically, we demonstrate the use of a well-known approach, Importance Sampling (IS): a reweighting of rewards generated by the behavior policy, $\pi_b$, such that they are equivalent to unbiased rewards from an evaluation policy $\pi_e$.

Let us begin by defining the IS procedure for POMDPs. Suppose we are given a trajectory $\tau \in \mathcal{T}_t$. We can express $P^e(\tau)$ using $P^b(\tau)$ as

$$\begin{aligned} P^e(\tau) &= P^e(u_0, z_0, a_0, \ldots, u_t, z_t, a_t) \\ &= \nu_0(u_0) \prod_{i=0}^{t-1} P^b(u_{i+1}|u_i, a_i) \prod_{i=0}^{t} P^b(z_i|u_i) \pi_e^{(i)}(a_i|h_i^o) \\ &= P^b(\tau) \prod_{i=0}^{t} \frac{\pi_e^{(i)}(a_i|h_i^o)}{\pi_b^{(i)}(a_i|u_i)}. \end{aligned}$$

We refer to $w_i = \frac{\pi_e^{(i)}(a_i|h_i^o)}{\pi_b^{(i)}(a_i|u_i)}$ as the importance weights. The importance weights allow us to evaluate $v(\pi_e)$ using the data generating process $P^b(\cdot)$ as

$$\text{IS}(\pi_e, w) := \mathbb{E}\left( R_L(\tau) \prod_{i=0}^{L} w_i \middle| \tau \sim \pi_b, u_0 \sim \nu_0 \right). \quad (1)$$

Note that $v(\pi_e) = \text{IS}(\pi_e, w)$. Unfortunately, the above requires the use of $\pi_b^{(i)}(a_i|u_i)$, which are unknown and *cannot* be estimated from data, as $u_i$ are unobserved under the POMDP model.

### Sufficient Condition for Importance Sampling

As Equation (1) does not resolve the off-policy problem, it remains an open question whether an IS procedure can be used in general POMDPs. Here, we give sufficient conditions for which a variant of IS can properly evaluate $\pi_e$. More specifically, we assume a POMDP which satisfies the following condition.

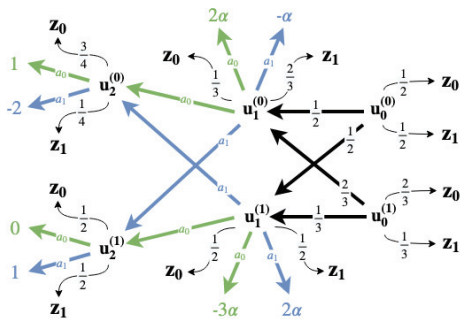



Figure 2: An example of a POMDP with 6 states and 2 observations for which importance sampling with importance weights $w_i = \frac{\pi_e^{(i)}(a_i|h_i^o)}{P^b(a_i|h_i^o)}$ is biased. Numbers on arrows correspond to probabilities. Arrows marked by $a_0, a_1$ correspond to rewards or transitions of these actions. Rewards depend on values of $\alpha > 0$. Initial state distribution is $\nu_0 = (\frac{1}{2}, \frac{1}{2})$.

**Assumption 1** (Sufficient Condition for IS)**.**

$$\exists f_L : \mathcal{T}_L^o \mapsto \mathbb{R} \; s.t. \; \forall \tau \in \mathcal{T}_L, R_L(\tau) = f_L(\tau^o).$$

In other words, this assumption states that the observed trajectory at time $L$ is a sufficient statistic of the true reward. Under Assumption 1 we can construct an IS variant as follows. Given a trajectory $\tau \in \mathcal{T}_L$ we have that

$$\begin{aligned} v(\pi_e) &= \sum_{\tau \in \mathcal{T}_L} R(\tau) P^e(\tau) \\ &= \sum_{u_0, \ldots, u_L} \sum_{\tau^o \in \mathcal{T}_L^o} f_L(\tau^o) P^e(\tau^o, u_0, \ldots, u_L) \\ &= \sum_{\tau^o \in \mathcal{T}_L^o} f_L(\tau^o) P^e(\tau^o). \end{aligned}$$

We can express $P^e(\tau^o)$ using $P^b(\tau^o)$ for any $\tau^o \in \mathcal{T}_L^o$ as

$$\begin{aligned} P^e(\tau^o) &= P^e(z_0, a_0, \ldots, z_t, a_t) \\ &= P^b(z_0) \prod_{i=0}^{L-1} P^b(z_{i+1}|\tau_i^o) \prod_{i=0}^{L} \pi_e^{(i)}(a_i|h_i^o) \\ &= P^b(\tau^o) \prod_{i=0}^{L} \frac{\pi_e^{(i)}(a_i|h_i^o)}{P^b(a_i|h_i^o)}. \end{aligned}$$

We can thus evaluate $v(\pi_e)$ using Equation (1) and importance weights $w_i = \frac{\pi_e^{(i)}(a_i|h_i^o)}{P^b(a_i|h_i^o)}$. While this estimate seems simple and intuitive, as we demonstrate next, arbitrarily large evaluation errors can occur using the above IS weights when Assumption 1 does not hold.

### Importance Sampling Error in POMDPs

In general POMDPs, if Assumption 1 does not hold, using the importance weights $w_i = \frac{\pi_e^{(i)}(a_i|h_i^o)}{P^b(a_i|h_i^o)}$ can result in large errors in the evaluation of $\pi_e$. This error is demonstrated by an example given in Figure 2. In this example, we assume an initial state distribution $\nu_0 = (\frac{1}{2}, \frac{1}{2})$. Given a

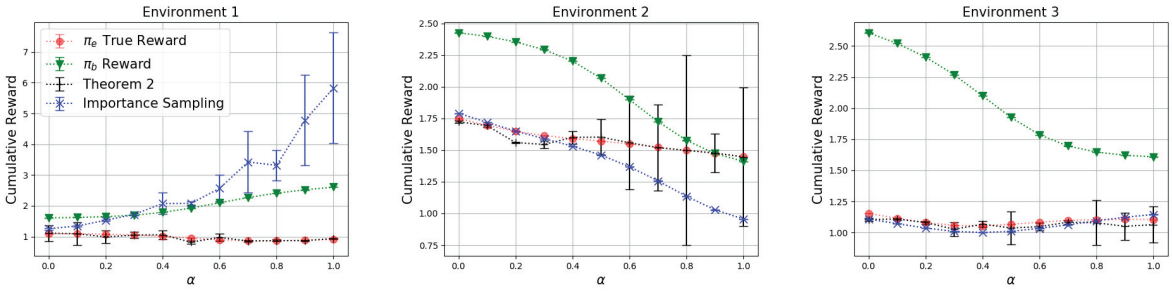



Figure 3: Comparison of cumulative reward approximation on three distinct synthetic environments. $\pi_b$ (green, triangles) and $\pi_e$ (red, circles) plots depict the true cumulative rewards of the behavior and evaluation policies, respectively. Ideally we would want the black "Theorem 2" curve and the blue IS curve to match the red curve of the true reward.

behavior policy $\pi_b(a_i|u_k^{(j)}) = \begin{cases} \frac{2}{3} & i \oplus j = 0 \\ \frac{1}{3} & i \oplus j = 1 \end{cases}$ for all $k$, and a stationary evaluation policy $\pi_e(a_i|z_j) = \begin{cases} \frac{2}{3} & i \oplus j = 0 \\ \frac{1}{3} & i \oplus j = 1 \end{cases}$, we have that $v(\pi_b) \approx 0.72\alpha + 0.26\gamma$ and $v(\pi_e) \approx -0.01\alpha + 0.14\gamma$. However, using $w_i = \frac{\pi_e(a_i|z_i)}{P^b(a_i|h_i^o)}$, IS evaluation yields $\text{IS}(\pi_e, w) \approx 0.62\alpha + 0.34\gamma$. This amounts to an error of $0.63\alpha + 0.2\gamma$ between the true policy evaluation and its importance sampling evaluation, which can be arbitrarily large. As an example, for $\alpha = 0.8\gamma$ we get that $v(\pi_b) = \text{IS}(\pi_e, w)$. Particularly, $\text{IS}(\pi_e, w) > v(\pi_b)$ for $\alpha > 0.8$. This is contrary to the fact that $v(\pi_b) > v(\pi_e)$ for all values of $\alpha > 0$.

Unlike the invertibility assumptions of Theorems 1 and 2, Assumption 1 is a strong assumption, that is unlikely to hold for almost any POMDP. At this point it is unclear if a more lenient assumption can assure an unbiased IS estimator for POMDPs, and we leave this as an open research question for future work. In the next section, we experiment with the results of Theorem 2 and the IS variant constructed in this section on a finite-sample dataset generated by a synthetic medical environment.

## 6 Experiments

The medical domain is known to be prone to many unobserved factors affecting both actions and rewards (Gottesman et al. 2018). As such, OPE in these domains requires adaptation to partially observable settings. In our experiments we construct a synthetic medical environment using a Decoupled POMDP model, as described next.

We denote $\sigma(x) = \frac{1}{1+e^{-x}}$. The environment consists of a patient's (observed) medical state $z$. Here $z$ changes according to the taken action by $P(z'|z,a) \propto \sigma(c^\top_{z,z',a}\phi_z(z))$, where $c_{z,z',a}$ are given weights, and $\phi_z(z)$ are the state features. We assume there exist unobserved variables $u = (u_{\text{mood}}, u_{\text{look}})$ relating to the doctors current mood, and how "good" the patient looks to her, respectively. In other words, we assume the doctor has an inner subjective ranking of the patient's look. Observation of the doctor's mood and inner ranking are modeled by $P(o|u) \propto \sigma(c^\top_{u,o}\phi_u(u))$, where $c_{u,o}$ are given weights, and $\phi_u(u)$ are the unobserved state features. Such observations could be based on the doctor's textual notes. Such notes, when processed through state of the art sentiment analysis algorithms (Qian et al. 2016) act as proxies to the doctor's mood and subjective assessment of the patient.

We model the doctor's mood changes according to the doctor's current mood, patient's look, and taken action $P(u'_{\text{mood}}|u,a) \propto \sigma(c^\top_{u,a,u'_{\text{mood}}}\phi_{\text{mood}}(u))$. Finally, the doctor's inner ranking of the patient's look is dependent on the patient's state $z$ and look $u_{\text{look}}$ by $P(u'_{\text{look}}|z', u_{\text{look}}) \propto \sigma(c^\top_{z',u_{\text{look}},u'_{\text{look}}}(\phi_{\text{look}}(u_{\text{look}}), \phi_z(z)))$.

We assume data generated by a confounded reward function and behavior policy

$$r(u,z,a) \propto \sigma((1-\alpha)(c^r_{z,a})^\top\phi_z(z) + \alpha(c^r_{u,a})^\top\phi_u(u)),$$
$$\pi_b(a|u,z) \propto \sigma((1-\alpha)(c^b_{z,a})^\top\phi_z(z) + \alpha(c^b_{u,a})^\top\phi_u(u)).$$

Here $\alpha \in [0,1]$ is a parameter which controls the "level of confoundedness". In other words, $\alpha$ is a measure of the intensity in which $\pi_b$ and $r$ depend on the unobserved state $u$, with $\alpha = 0$ corresponding to no unobserved confounding.

The observed state space $\mathcal{Z}$, unobserved state space $\mathcal{U}$, and observation space $\mathcal{O}$ were composed of two binary features each. We run the experiment in three environments, corresponding to different settings of the vectors $c$ meant to illustrate different behaviors of our methods. Ten million trajectories were sampled from the policy $\pi_b$ over a horizon of 4 time steps for each environment. Figure 3 depicts the cumulative reward of $\pi_e$, $\pi_b$, and their corresponding estimates according to Theorem 2 and the IS weights $w_i^k = \frac{\pi_e^{(i)}(a_i|h_i^o)}{P^b(a_i|h_i^o)}$, for different values of $\alpha$.

Environment 1 illustrates a typical result from the generative process above, where the vectors $c$ were sampled from a Gaussian. It is clear that IS estimates increase in bias with $\alpha$, whereas Theorem 2's estimate remains unchanged. Moreover, for values of $\alpha > 0.3$, IS suggests that $\pi_e$ is superior to $\pi_b$. This implies that potentially arbitrarily bad policies could be learned by an off-policy RL algorithm. Environments 2 and 3 are atypical, and were found through deliberate search, to illustrate situations in which our estimation procedure does not clearly outperform IS. In Environment 2, for large values of $\alpha$, variance increases, due to near non-invertibility of the conditional probability matrices. In Environment 3, despite confounding, IS remains unbiased.

## 7 Related Work

**POMDPs:** Uncertainty is a key feature in real world applications. POMDPs provide a principled general framework for planning under uncertainty (Spaan 2012; Williams and Young 2007). POMDPs model aspects such as the stochastic effects of actions, incomplete information and noisy observations over the environment. POMDPs are known to be notoriously hard to approximate (Madani, Hanks, and Condon 1999; Papadimitriou and Tsitsiklis 1987). Their intractability is mainly due to the "curse of dimensionality" for which complexity grows exponentially with the cardinality of the unobserved state space. As this work considers offline evaluation under uncertainty, the unobserved state is treated as a confounding element for policy evaluation.

To the best of our knowledge our work provides the first OPE results for POMDPs. Nevertheless, there has been considerable work on learning in POMDPs, where, similar to our setting, the agent does not gain access to the POMDP model. Even-Dar, Kakade, and Mansour (2005) implement an approximate reset strategy based on a random walk of the agent, effectively resetting the agent's belief state. Hausknecht and Stone (2015) tackle the learning problem by adding recurrency to Deep Q-Learning, allowing the Q-network to estimate the underlying system state, narrowing the gap between $Q_\theta(o, a)$ and $Q_\theta(u, a)$. Nevertheless, work on learning in POMDPs greatly differs from OPE for POMDPs, as the former offer online solutions based on interactive environments, whereas the latter uses batch data generated by an *unknown* and *unregulated* behavior policy.

**Off-Policy Evaluation (OPE):** Contemporary OPE methods can be partitioned into three classes: (1) direct methods (DM) (Precup 2000; Munos et al. 2016; Le, Voloshin, and Yue 2019; Jiang and Li 2015), which aim to fit the value of a policy directly, (2) inverse propensity score (IPS) methods, also known as importance sampling (IS) methods (Liu et al. 2018; Dudík, Langford, and Li 2011; Jiang and Li 2015), and (3) Doubly-Robust methods (DRM) (Jiang and Li 2015; Thomas and Brunskill 2016; Kallus and Uehara 2019), which combine IS methods with an estimate of the action-value function, typically supplied by a DM. These algorithms were designed for bandits, and later generalized to RL. Nevertheless, existing methods assume full observability of the underlying state. They become dubious when part of the data generating process is unobserved or unknown.

**Causal Inference:** A major focus of work in causal inference is how to estimate, in an offline model, the effects of actions without fully observing the covariates which lead to the action (Pearl 2009; Spirtes et al. 2000). Much of the work in this field focuses on static settings, with some more recent work also tackling the bandit setting (Bareinboim, Forney, and Pearl 2015; Forney, Pearl, and Bareinboim 2017; Ramoly, Bouzeghoub, and Finance 2017; Sen et al. 2016). Sufficient "sequential ignorability" conditions (no hidden confounding) and methods for OPE of causal effects under dynamic policies are given by (Murphy et al. 2001; Hernán et al. 2006; Hernán and Robins 2019).

Recently, there has been growing interest in handling unobserved confounders in the context of MDPs. Zhang and Bareinboim (2016) consider a class of counterfactual policies that incorporate a notion they call "intuition", by using observed actions as input to an RL agent. Their confounding model is a special case of our proposed Decoupled POMDP model in which confounding factors are independent of each other. Lu, Schölkopf, and Hernández-Lobato (2018) propose a latent variable model for "deconfounding reinforcement learning". They extend the work of Louizos et al. (2017) by positing a deep latent variable model with a single unobserved confounder that governs a trajectory, deriving a variational lower bound on the likelihood and training a model with variational inference. Their causal model does not take into account dynamics of unobserved confounders. Oberst and Sontag (2019) also look at off-policy evaluation in POMDPs, though unlike this work they assume that the unobserved state does *not* directly affect the observed behavior-policy actions. Their work focuses on counterfactuals: what would have happened in a specific trajectory under a different policy, had all the other variables, including the random noise variables, been the same. This is a difficult task, lying on the third rung of Pearl's causal hierarchy, which we restate in the supplementary material. (Pearl 2018). Our task is on the second rung of the hierarchy: we wish to know the effect of intervening on the world and acting differently, using a policy $\pi_e$. Oberst and Sontag (2019) therefore requires more stringent assumptions than ours on the structure of the causal model, namely an extension of outcome monotonicity.

Our work specifically extends the work of Miao, Geng, and Tchetgen Tchetgen (2018), which rests on the measurement of two independent proxy variables in a bandit setting. Our results generalizes their identification strategy through the independence structure that is inherent to POMDPs and Decoupled POMDPs, where past and future are independent conditioned on the unobserved confounder at time $t$.

## 8 Conclusion and Future Work

Off-policy evaluation of sequential decisions is a fundamentally hard problem, especially when it is done under partial observability of the state. Unknown states produce bias through factors that affect both observed actions and rewards. This paper offers one approach to tackle this problem in POMDPs and Decoupled POMDPs.

While the expressiveness of POMDPs is useful in many cases, it also comes with a substantial increase in complexity. Yet, one may not necessarily require the complete general framework to model complex problems. This paper takes a step towards an alternative model, Decoupled POMDP, for which unobserved factors are isolated, reducing OPE complexity, while maintaining the same expressive power as POMDPs. We note that Decoupled POMDPs may also benefit general purpose RL algorithms in partially observable environments.

In this work we experimented with a tabular environment. As future work, one may scale up to practical domains using latent space embeddings of the generalized weight matrices, as well as sophisticated sampling techniques that may reduce variance in approximation.

## 9 Acknowledgments

We thank Michael Oberst, Moshe Tennenholtz, and the anonymous reviewers for their fruitful comments that greatly improved this paper. Research was conducted under ISF grant number 1380/16.

## References

Bareinboim, E.; Forney, A.; and Pearl, J. 2015. Bandits with unobserved confounders: A causal approach. In *Advances in Neural Information Processing Systems*, 1342–1350.

Bordenave, C.; Caputo, P.; and Chafaï, D. 2012. Circular law theorem for random Markov matrices. *Probability Theory and Related Fields* 152(3-4):751–779.

Dann, C.; Neumann, G.; and Peters, J. 2014. Policy evaluation with temporal differences: A survey and comparison. *The Journal of Machine Learning Research* 15(1):809–883.

Dudík, M.; Langford, J.; and Li, L. 2011. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*.

Even-Dar, E.; Kakade, S. M.; and Mansour, Y. 2005. Reinforcement learning in POMDPs without resets. In *Proceedings of the 19th international joint conference on Artificial intelligence*, 690–695. Morgan Kaufmann Publishers Inc.

Forney, A.; Pearl, J.; and Bareinboim, E. 2017. Counterfactual data-fusion for online reinforcement learners. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1156–1164. JMLR. org.

Gottesman, O.; Johansson, F.; Meier, J.; Dent, J.; Lee, D.; Srinivasan, S.; Zhang, L.; Ding, Y.; Wihl, D.; Peng, X.; et al. 2018. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*.

Hausknecht, M., and Stone, P. 2015. Deep recurrent q-learning for partially observable MDPs. In *2015 AAAI Fall Symposium Series*.

Hernán, M. A., and Robins, J. M. 2019. *Causal Inference*. Chapman & Hall/CRC.

Hernán, M. A.; Lanoy, E.; Costagliola, D.; and Robins, J. M. 2006. Comparison of dynamic treatment regimes via inverse probability weighting. *Basic & clinical pharmacology & toxicology* 98(3):237–242.

Jiang, N., and Li, L. 2015. Doubly robust off-policy value evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*.

Kallus, N., and Uehara, M. 2019. Double reinforcement learning for efficient off-policy evaluation in Markov decision processes. *arXiv preprint arXiv:1908.08526*.

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

Le, H. M.; Voloshin, C.; and Yue, Y. 2019. Batch policy learning under constraints. *arXiv preprint arXiv:1903.08738*.

Liu, Q.; Li, L.; Tang, Z.; and Zhou, D. 2018. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, 5356–5366.

Louizos, C.; Shalit, U.; Mooij, J. M.; Sontag, D.; Zemel, R.; and Welling, M. 2017. Causal effect inference with deep latent-variable models. In *Advances in Neural Information Processing Systems*, 6446–6456.

Lu, C.; Schölkopf, B.; and Hernández-Lobato, J. M. 2018. Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576*.

Madani, O.; Hanks, S.; and Condon, A. 1999. On the undecidability of probabilistic planning and infinite-horizon partially observable Markov decision problems. In *AAAI/IAAI*, 541–548.

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

Munos, R.; Stepleton, T.; Harutyunyan, A.; and Bellemare, M. 2016. Safe and efficient off-policy reinforcement learning. In *Advances in Neural Information Processing Systems*, 1054–1062.

Murphy, S. A.; van der Laan, M. J.; Robins, J. M.; and Group, C. P. P. R. 2001. Marginal mean models for dynamic regimes. *Journal of the American Statistical Association* 96(456):1410–1423.

Oberst, M., and Sontag, D. 2019. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In *International Conference on Machine Learning*, 4881–4890.

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

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. New York, NY, USA: Cambridge University Press, 2nd edition.

Pearl, J. 2018. Theoretical impediments to machine learning with seven sparks from the causal revolution. *arXiv preprint arXiv:1801.04016*.

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

Puterman, M. L. 1994. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

Qian, Q.; Huang, M.; Lei, J.; and Zhu, X. 2016. Linguistically regularized LSTMs for sentiment classification. *arXiv preprint arXiv:1611.03949*.

Ramoly, N.; Bouzeghoub, A.; and Finance, B. 2017. A causal multi-armed bandit approach for domestic robots' failure avoidance. In *International Conference on Neural Information Processing*, 90–99. Springer.

Romanovskii, I. 1965. Existence of an optimal stationary policy in a Markov decision process. *Theory of Probability & Its Applications* 10(1):120–122.

Sen, R.; Shanmugam, K.; Kocaoglu, M.; Dimakis, A. G.; and Shakkottai, S. 2016. Contextual bandits with latent confounders: An NMF approach. *arXiv preprint arXiv:1606.00119*.

Spaan, M. T. 2012. Partially observable Markov decision processes. In *Reinforcement Learning*. Springer. 387–414.

Spirtes, P.; Glymour, C. N.; Scheines, R.; Heckerman, D.; Meek, C.; Cooper, G.; and Richardson, T. 2000. *Causation, prediction, and search*. MIT press.

Sutton, R. S., and Barto, A. G. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Thomas, P., and Brunskill, E. 2016. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, 2139–2148.

Williams, J. D., and Young, S. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language* 21(2):393–422.

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