Causality and Batch Reinforcement Learning: Complementary Approaches To Planning In Unknown Domains

James Bannon
Tao Li
Wenbo Song
Brad Windsor

Abstract

Reinforcement learning algorithms have had tremendous successes in online learning settings. However, these successes have relied on low-stakes interactions between the algorithm/agent and its environment. In many settings where RL could be of use, such as health care and autonomous driving, the mistakes made by most online RL algorithms during early training come with unacceptable costs. These settings require developing reinforcement learning algorithms that can operate in the so-called *batch* setting, where the algorithms must learn from set of data that is fixed, finite, and generated from some (possibly unknown) policy. Evaluating policies different from the one that collected the data is called *off-policy evaluation*, and naturally poses *counter-factual* questions. In this project we show how off-policy evaluation and the estimation of average treatment effects in causal inference are two approaches to the same problem, and compare recent progress in these two areas.

1. Introduction

The Limits of Limitless Data Reinforcement learning (RL) distinguishes itself from supervised and unsupervised learning by being interactive Sutton and Barto (2018). In the typical RL setting, an agent interacts with their environment in an ongoing way with the goal of learning a *policy* that is optimal in some sense. Theoretical analyses of these algorithms often focus on long term optimality, either the convergence to a truly optimal policy or a bound on the long term *regret* of the algorithm. Indeed, high profile RL successes in playing games like backgammon Tesauro (1995), Go Silver et al. (2016), Quake IIII Jaderberg et al. (2019), and StarCraft II Vinyals et al. (2019) have been in this setting.

What these settings have in common is that the only cost to collecting additional data is computational. Put differently, early mistakes made by the algorithm/agent before (asymptotic) convergence only require that the agent be restarted and have no consequences in the physical world. Many of the aforementioned successes leverage this lack of stake by running for several million iterations.

This dependence on the ability to acquire an arbitrarily large amount of data proves to be a strong limitation when attempting to apply these algorithms to certain real-world scenarios. In healthcare, autonomous driving, or automated plant control, the costs of making mistakes are unacceptably high, and collecting millions of data points is infeasible. On the contrary, reinforcement learning algorithms must proceed from a fixed set of data

$$D = \{(s_t, a_t, s_{t+1}, r_t)\}_{t=1}^{T}$$

which may be generated from collection of (possibly unknown) policies π_t where $a_t = \pi_t(s_t)$. The challenges to batch RL algorithms come primarily from the fact that they must do *off-policy evaluation*, that is they need to estimate V^π of some policy π that may differ arbitrarily from the policies used to collect D. In estimating the value of a policy different to the one that generated the data, an implicitly counterfactual question is asked: "what would the value of the total reward been the algorithm had performed a'_t instead of a?"

Counterfactual questions are foundational to the field of causal inference. As such, off-policy evaluation in reinforcement learning looks a lot like counter-factual inference in statistics.

In this project report we first discuss causality and causal inference. We then discuss its relationship to causal RL and off-policy evaluation. Finally we argue that these two areas can benefit from mutual dialogue.

2. Causality

Causality is best understood as the science of predicting the results of *interventions*. This sets it apart from typical statistics and machine learning problems where the chief concern is make predictions based on *observations*. In the typical statistical learning scheme the goal is to design an algorithm to use a labeled data set $\{(X_i, Y_i)\}_{i=1}^n$ to learn a function $f : X \to Y$ which can then be used to make predictions $f(X)$ on new data items X. Causality cares about designing algorithms predicting the Y when X is actively manipulated.

The distinction arises from the fact that observing X does not include latent influences that lead to that observation. Consider the following example: data $\{(X, Y)\}$ is collected on a population where $X \in \{0, 1\}$ represents a person taking a specific multi-vitamin and $Y \in \mathbb{R}$ is their average daily cortisol levels. For the i'th person, predicting Y_i from observing X_i ignores the processes that lead to observing that X_i occurred. If, for example, people with higher personal wealth have no benefit from the multivitamin but *more* likely to have a meditation app on their phone, then the observational data will be *confounded*. If the data set is collected from a mostly wealthy population, then it will appear that there is a positive association between taking the vitamin and lowered cortisol level. In fact if new samples are drawn from that same population it is possible to predict cortisol levels with a high degree of accuracy given just the fact that they take a multivitamin, but this won't provide any information about *how* this influence occurs. Put differently, just because we can *predict* the cortisol levels from the fact that someone takes a multivitamin, doesn't mean we can tell what will happen to someone's cortisol if we *force them to take the multivitamin.*

In the causality literature, which spans statistics, computer science, and mathematics, among other areas, there are two distinct but symbiotic kinds causal questions. The first kind of question is **causal discovery**, that is, given a collection of variables $X_1, \ldots, X_k$, attempting to learn causal relationships between them. This depends on a chosen semantics of causal relationships and comes with some computational difficulties. For example, bayesian networks, one way of representing causal relationships, are NP-hard to learn Chickering et al. (2004). That's said, many heuristic approaches exist in the literature, including several neural-network based ones. **BRAD MAYBE PUT SOME STUFF HERE?**

The second kind of question is **causal inference.** In causal inference hypothesized causal relationships between covariates is assumed and the subsequent task is to infer *the*

degree to which one has a causal effect on another. For the rest of this project, we abandon the question of causal discovery and focus on causal inference as it is applied to batch reinforcement learning. In particular, as most RL algorithms assume an underlying MDP, a causal relationship between variables is already to an extent implicit and so the task is

Before we can discuss inference rigorously, we need to have a formal representation of causal relationships, which we turn to now.

2.1 Representing Causal Relationships & Interventions

Causal inference has a long history across many disciplines. In statistics, the *potential outcomes framework* has dominated the literature Rubin (2005); Splawa-Neyman et al. (1990). Two similar approaches to modeling causality are the structural equation models (SEM) approach in economics Duncan (2014); Goldberger (1973) and the graphical models framework from Bayesian inference Neapolitan et al. (2004).

Definition 1 (Structural Causal Model) *A structure causal model M is a tuple $M = (V, U, F, Pa(\cdot), P(U))$ where*

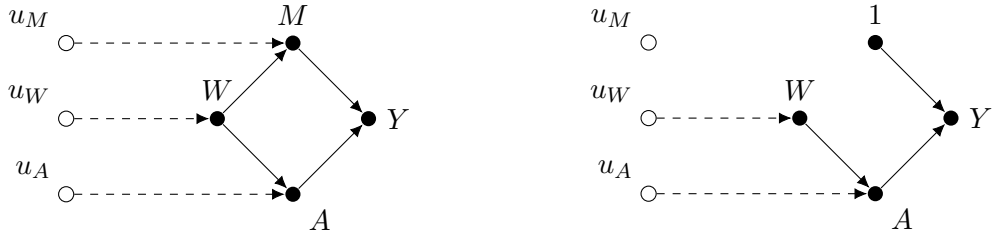
- *V is a set of observed variables (sometimes called* **endogenous** *variables)*
- *U is a set of unobserved random variables (sometimes called* **exogenous** *variables.*
- *$Pa : S \subset V \cup U \to (V \cup U) - S$ is the* **parent function**.
- *$P(U)$ is a probability distribution over the exogenous variables.*
- *F is a family of functions f_V mapping $f_V : Pa(V) \to range(V)$*

Intuitively the SCM puts all the stochasticity into the exogenous variables U and then makes the observed variables V deterministic functions of their exogenous and endogenous parents. Note that Theorem 1 abstracts away any graphical representation but has a natural expression as a directed graph. This graph, however, is not necessarily acyclic. To continue the example from above, let M, W, A be the observed variables for a person taking the **m**ulti-vitamin, being **w**ealthy, and having the meditation **a**pp on their phone, respectively, and let u_i be Bernoulli with parameter p_i for $i \in \{M, W, A\}$. The fully written SCM would be:

$$\begin{aligned} U_W &\sim \text{Ber}(p_W) & W &= U_W \\ U_A &\sim \text{Ber}(p_A) & A &= \max\{W, U_A\} \\ U_M &\sim \text{Ber}(p_M) & M &= \max\{W, U_M\} \\ & & Y &= \mathcal{N}(-3 * A - 1 * M, 1) \end{aligned}$$

Intervening by setting M to 1 corresponds to changing the equation to M Suppose $Y \sim \mathcal{N}(-3 * A - M, 1)$ where $\mathcal{N}$ represents a normal density.

Individual treatment effects



(a) The full causal model

(b) Intervening on M by setting it to 1

Figure 1

Truncated factorization: Any markov model we have:

$$P(v_1, \ldots, v_k | \text{do}(x_0)) = \prod_{i:V_i \in X} P(v_i | Pa(v_i)\big|_{x=x_0} \tag{1}$$

where $P(v_i|pa(v_i))$ are the pre-intevention conditional probabilities

3. A Warmup on Bandit Problems

Contextual bandit problems **???** provide a natural middle ground between MDP-based reinforcement learning,

Causal contexts: one option is to have the bandits existing in a causal context **?**.

4. Off-Policy Policy Evaluation: Causal Approaches

As we have briefly discussed in the introduction, offline reinforcement learning rests on the same counterfactual thinking as causal inference does (counterfactual inference). With missing data or limited access to data, both of them deal try to extract some information from historical data or observations and apply it to counterfactual scenarios, where it is too costly to launch experiments and collect data or too controversial even unethical to rollout online exploration (e.g. healthcare applications). In this section, we focus on causality based approaches in solving off-policy policy evaluation (OPPE) problem, which is the key challenge in off-line reinforcement learning and we shall show that OPPE can either be reformulated as a counterfactual inference problem in a model-free fashion or can be solved in a model-based manner where the model is constructed by counterfactual thinking. To be more specific, in this section, we first give a quick review about Markov Decision Process (MDP) Bellman (1957) and introduce off-policy policy evaluation problem. Then, we move to recent advances on OPPE based on causal inference and we categorize those methods into two parts: **model-free** and **model-based** ones, showing that these causality-based approaches bridge causal inference and off-line reinforcement learning. Finally, we make a comparison between causality-based and existing methods for OPPE, where we illustrate the advantages of incorporating causality into OPPE.

4.1 Offline Reinforcement Learning

In this subsection, we provide a brief overview of off-line reinforcement learning. We use the notational standard MDPNv1 Thomas and Okal (2015) for better presenting our introductions. Generally, in a reinforcement learning problem, the task of the agent who are supposed to make decisions sequentially in a dynamic environment, is to find an optimal policy maximizing the expected long term return, based on rewards received at each time step. The challenge of reinforcement learning is that the agent does not have complete information about the dynamic environment and tries to identify high-reward behavior patterns by interacting with the unknown environment. This sequential decision-making process under some dynamic environment is modeled by Markov Decision Process Sutton and Barto (2018), denoted by a tuple $(\mathcal{S}, \mathcal{A}, P, \mathcal{R}, R, d_0, \gamma)$, where

- We use $t \in \mathbb{N}_{\geq 0}$ to denotes the time step, where $\mathbb{N}_{\geq 0}$ denotes the natural numbers *including zero.*
- $\mathcal{S}$ is the state space that the agent can be in, and is called the *state set.* We note that the state of the environment at time t is a random variable denoted by S_t. We will typically use s to denote an element of the state set.
- $\mathcal{A}$ is the action space, i.e, the set of possible actions the agent can perform. The action at time t is denoted by A_t, while a denotes an element of the action set.
- $\mathcal{R}$ is the set of possible rewards, defined as $\mathcal{R} \subseteq \mathbb{R} \cup \{-\infty, \infty\}$. Additionally, instantaneous reward at time t is $r(s_t, a_t)$, where s_t, a_t, as defined above are the current state and action respectively.
- $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \longmapsto [0, 1]$ is called the *transition function.* For all $(s, a, s', t) \in \mathcal{S} \times \mathcal{A} \times \mathcal{S} \times \mathbb{N}_{\geq 0}$, let $P(s, a, s') \coloneqq \Pr(S_{t+1} = s' \mid S_t = s, A_t = a)$. In other words, P characterizes the distribution over states at time $t+1$ given the state and action at time t. We note that this distribution is only conditional on the last state and action and is independent of the previous history, indicating that the sate transition is *markovian.*
- d_0 is the initial distribution of states defined as $d_0 : \mathcal{S} \longmapsto [0, 1]$ and γ is the discount factor defined as $\gamma \in [0, 1)$.

As we mentioned above, the goal of the agent is to find a policy $\pi : \mathcal{S} \times \mathcal{A} \longmapsto [0, 1]$ defined as a probability distribution over $\mathcal{S} \times \mathcal{A}$, such that by following this policy, the agent maximize the expected return. For the sake of simplicity, we here consider finite horizon MDP, where the trajectory in this setting is of finite horizon length, e.g., $\tau = (s_0, a_0, \ldots, s_H)$ is a trajectory, a sequence of states and actions up to s_H the terminal state. We note that if a trajectory τ is generated by following a policy μ, then the distribution of τ can be explicitly computed as follows, given the transition probability $P(s, a, s')$: $p_\mu = d_0(s_0) \prod_{t=0}^{H-1} [P(s_t, a_t, s_{t+1}) \mu(s_t, a_t)]$, where d_0 is the initial state distribution. Therefore, the ultimate goal for one agent is to find a policy π such that the expected return of this policy $\mathbb{E}_{\tau \sim p_\pi}\{\sum_{i=0}^{H} r(s_t, a_t)\}$ is maximized. Generally we treat the expected return as the quality of the policy and as we shall see later, the estimation of the policy quality is of vital importance in reinforcement learning. Besides this expected return, there are also other

related notations and quantities. For example, another way to evaluate a policy is to look at the conditional expected return that is conditional on the initial state s_0, which is referred to as the value function $V(s) := \mathbb{E}_{p_\pi}\{\sum_{i=0}^{H} r(s_t, a_t)|s_0 = s\}$. Also, different MDP models M admit different transition kernels, hence the same policy may induce different distributions over the trajectory, and to make it more clear, we usually specify the MDP model in the notation $p_{M,\mu}$. All these subtle differences in notations and definitions are skipped for now and shall be detailed in the sequel whenever necessary.

Powered by advances in deep learning and computing power, we have seen many great successes in reinforcement learning, which has been viewed as a very general learning paradigm that can model a wide range of problems, such as games Mnih et al. (2015); Silver et al. (2016) , robotics Kober et al. (2013), autonomous driving Sallab et al. (2017), healthcare Komorowski et al. (2018), and many others. For many reinforcement learning practice, such as the examples we mentioned above, as argued in Li (2019), the quality of a policy is often measured by the average reward received if the policy is followed by the agent to select actions. That is, many successful reinforcement learning applications rely on on-policy data, collected through online learning. However, before reinforcement learning can be widely adopted in real-life applications, we must address the limitation of online learning, where the agent can evaluate and learn the policy through online exploration, because for most real-life problems like autonomous driving and medical treatment, running a new policy in the actual environment can be costly, risky and/or unethical.

Offline reinforcement learning aims at learning a new policy based on only historical observations and additional interactions with the unknown environment are not permitted Lange et al. (2012). The inability of receiving feedback from the environment avoids potential costs and safety issues but also poses a great challenge to the learning process. As observed in Fujimoto et al. (2019); Wu et al. (2019), policy evaluation is a difficult problem in offline reinforcement learning, since the learned value functions, such as Q-values are often not a good indicator of the performance and if we cannot evaluate the policy properly, then we cannot improve the policy. Hence, to evaluate the learned policy from the data generated by the behavioral policy, we have to leverage off-policy policy evaluation, as we shall introduce in the following.

Off-policy evaluation, or more broadly speaking, policy evaluation is about measuring the quality of a given policy, and this measurement rests on the sample trajectories either generated by the policy to be evaluated (on-policy) or by some different policy or policies (off-policy). Generally, off-policy policy evaluation is more challenging than on-policy policy evaluation, since there is a distributional mismatch between the policy we are interested in and the policy that is actually implemented in the MDP model for generating those sample trajectories. In the sequel, we refer to the interested policy as the target policy, denoted by π while to the data-generating policy as the behavioral policy μ. To see this mismatch, we only need to realize that sample trajectories are in fact a series of state-action pairs sampled from the trajectory distribution. For example, for a MDP model M, a given behavioral policy μ, and a sample trajectory of $H+1$ states and actions, $\tau = (s_0, a_0, \ldots, s_H)$, then the distribution of τ as $p_{M,\mu} = d_0(s_0)\prod_{t=0}^{H-1}[P(s_t, a_t, s_{t+1})\mu(s_t, a_t)]$, where d_0, as introduced at the beginning of this section, is the initial state distribution. Obviously, for the target policy $\pi \neq \mu$, we have the mismatch $p_{M,\mu} \neq p_{M,\pi}$, meaning that this sample trajectory τ cannot be directly applied to estimating the expected return $\mathbb{E}_{\tau\sim p_{M,\pi}}\{\sum_{i=0}^{H} r(s_t, a_t)\}$. In

order to address this distributional mismatch, many approaches have been proposed, we mainly categorize these methods into three parts: **model-based**, **model-free**, and **hybrid** ones, which are present in the following. Here, we merely give a bird's eye view of existing off-policy evaluation techniques and detailed discussions are included in the last subsection, where we look into the difference between them and those causality-based OPPE approaches.

- **Model-based**: This type of methods circumvent this mismatch by first learning a MDP model from the observations (sample trajectories) and then use the learned model to evaluate the target policy. Specifically, model-based methods focus on constructing an approximation of the reward function $r(s, a)$ and/or the transition kernel $P(s, a, s')$. For example, both approximate models in Jiang and Li (2016); Thomas and Brunskill (2016) rely on a reward function estimate of the underlying MDP model. For model estimation, the simplest method is probably the count-based estimation for discrete MDPs, where the reward function and the transition kernel are approximated based on observed state-action-reward tuples, leading to consistent maximum-likelihood estimators of the model. However, this method has no or limited generalization Paduraru (2013). If no samples are available for some state-action pair, then the associated estimation is not defined. Equipped with generalization, another common approach is regression estimator Paduraru (2013); Mannor et al. (2007) for continuous cases, which is essentially a representation learning for the MDP and its performance heavily depends on the selection of representation class and the design of loss functions. However, as pointed out in Farajtabar et al. (2018), there are two major problems with these model-based approaches: (1) Its bias cannot be easily quantified, since in general it is difficult to quantify the approximation error of a function class, and (2) It is not clear how to choose the loss function for model learning. The first problem is more about a choice of representations and causality cannot help with that but as we have shown in the last subsection. For the second one, we demonstrate that causal inference can bring a more reasonable loss function for model learning and we shall compare model-based estimators with causality-based model learning at the end of this section, where we show the advantage of counterfactual thinking during the model learning process.

- **Model-free**: Different from model-based methods, approaches we are introducing here, as its name suggests, do not rely on the underlying MDP model. Most model-free methods Guo et al. (2017); Precup et al. (2000); Thomas et al. (2015) aims directly at the distributional mismatch by importance sampling Kahn and Marshall (1953). As a technique in computational statistics, Importance Sampling (IS) computes an expected value under some distribution of interest by weighting the samples generated from some other distribution, which exactly fits the OPPE formulation. As we have introduced above, a sample trajectory $\tau = (s_0, a_0, \ldots, s_H)$ with a distrobution $p_{M,\mu}$, the empirical return $\sum_{t=0}^{H-1} r_t$ gives an unbiased estimate of the expected return $\mathbb{E}_{\tau\sim p_{M,\mu}}\{\sum_{i=0}^{H} r(s_t, a_t)\}$. If we define the per-step importance ratio as $\rho_t := \pi(s_t, a_t)/\mu(s_t, a_t)$ and the cumulative importance ration as $\rho_{0:t} := \prod_{k=0}^{t} \rho_k$, then the basic (trajectory-wise) IS estimator,

and an improved step-wise version are given as follows:

$$V_{\text{IS}} := \rho_{1:H} \cdot \left(\sum_{t=1}^{H} \gamma^{t-1} r_t \right), V_{\text{step-IS}} := \sum_{t=1}^{H} \gamma^{t-1} \rho_{1:t} r_t,$$

which is based on the fact that $p_{M,\pi} = d_0(s_0) \prod_{t=0}^{H-1} [P(s_t, a_t, s_{t+1}) \mu(s_t, a_t) \rho_t]$. For more details and variants such as weighted importance sampling, we refer readers to Precup et al. (2000); Thomas et al. (2015). The Advantage of model-free methods over model-based ones is obvious. Model-free methods provide a consistent estimator with low bias, whereas model-based ones often suffer from higher bias, if the estimated model is a poor approximation. However, IS-based approach is far from perfect and there are two major issues with it: 1) typically, this approach tends to have high variance, especially when the target policy is deterministic. 2) It requires absolute continuity assumption regarding the polices. That is, for all state-action pairs $(s, a) \in \mathcal{S} \times \mathcal{A}$, if $\mu(s, a) = 0$ then $\pi(s, a) = 0$. This assumption requires that all state-action pairs (s, a) produce by the target policy must have been observed in the sample trajectories, otherwise the importance ration ρ_t cannot be defined. To sum up, there is no generalization in IS approach, meaning that we are unable to deal with the data we haven't seen before and in the following we shall see how causality-based model-free methods tackle these issues with counterfactual inference, enabling generalization in off-policy evaluation when facing unobserved data.

- **Hybrid**: There has been a growing interest in combining the two approaches for constructing unbiased estimators with low variances. The first attempt in this direction can be dated back to Dudík et al. (2011), where a hybrid estimator named *doubly robust* (DR) estimator, combining both model estimate and importance sampling, is proposed for policy evaluation in contextual bandits. This DR estimator admits the following form: $V_{\text{DR}} := \widehat{V}(s) + \rho(r - \hat{r}(s, a))$, where $\hat{r}$ is a good estimate of the reward function, $\widehat{V}$ is the expected reward under the policy π, defined as $\widehat{V}(s) := \mathbb{E}_\pi\{\hat{r}(s, a)\}$ and ρ is the importance ratio. Simply put, we can treat this doubly robust estimator DR = model estimate + importance sampling. This idea is further developed in Jiang and Li (2016) and authors extend the DR for contextual bandit to a DR estimator for sequential decision-making. By decomposing the whole trajectory into state-action-reward tuples, off-policy evaluation in reinforcement learning can be viewed as estimation in multiple contextual bandits problems, where s_t is the state, a_t is the action, and the observed return is defined recursively for adding the temporal relations in MDP to the rewards in the contextual bandit setting. Built upon this DR estimator, many variants have been proposed recently. In Farajtabar et al. (2018), for the model-based part, a loss function (a weighted mean square error) is proposed for reducing the variance, yielding a better model estimation, which minimize the variance of the DR estimator. On the other hand, a novel approach for blending model-based and model-free parts is considered in Thomas and Brunskill (2016). In this work, model-based and model-free estimators are not applied to the sample trajectory at the same time, instead, authors first partition the whole trajectory into to parts, where the first half is handled by importance sampling and the rest by a model-based estimator. Naturally, different partitions lead to different estimates, blending model-free and model-based estimates

together differently and the new estimator in this paper returns a weighted average of these blending estimates so as to minimize the mean square error.

4.2 Causality-based Off-Policy Policy Evaluation

The idea behind causality-based off-policy policy evaluation is that the target policy is treated as a kind of intervention, comprising counterfactual actions that are different from the those in behavioral policy. Therefore, in order to evaluate the target policy based on observational data generated by the behavioral policy, we have to focus on the difference between the two policies or more specifically, the counterfactual outcomes. That is, *what would have happened, had we applied those actions from the target policy.* In other words, with this counterfactual thinking, OPPE problem reduces to counterfactual inference problem and in this subsection, we shall present two approaches of leveraging causal inference in solving OPPE problem.

4.2.1 Model-free Approach: Counterfactually-guided Policy Evaluation

The first approach does not rely on the underlying **MDP models**, and on the contrary, it casts MDPs into structrual causal models (SCM) for counterfactual inference, where actions in the target policy are viewed as interventions. We first introduce the generic approach of counterfactual inference then we move to its applications in OPPE, as studied in Buesing et al. (2018); Oberst and Sontag (2019).

Given a structural causal model, a counterfactual query is defined as a triple $(\hat{x}_0, I, X_q)$ of observations $\hat{x}_o$ of some variables X_o, an intervention I and query variables Xq. From the observations $\hat{x}_o$, we try to find out what the distribution of X_q is after applying interventions I to the given SCM. Generally, Counterfactual inference is performed as follows. We first extract some information about the unobserved confounders U from the observations $\hat{x}_o$, for example, we estimate the conditional distribution $p(U|\hat{x}_0)$. Then, with the information about confounders, we apply the interventions I and we compute $p^{\mathrm{do}(I)|\hat{x}_o}(x_q)$ the distribution of the query variables X_q and this distribution answers the counterfactual question *what would have happened, had we applied the interventions.*

On the other hand, as we discussed before, OPPE is also about answering a counterfactual question *what would have happened, had we applied those actions from the target policy.* Hence, if we can represent any given MDP models or more broadly POMDP models by an SCM over trajectories Buesing et al. (2018), and we treat those trajectories generated by the behavioral policy as observations, actions from the target policy as interventions, then we are also able to identify the expected return under the target policy using counterfactual inference, where the value function becomes the query variable. This one-to-one correspondence between counterfactual inference and OPPE, as shown in the Table1, is built upon structural models that capture the underlying cause-effect relationship between variables, which has been the foundation for some research works on causality-based decision-making, such as multi-armed bandits problem Bareinboim et al. (2015); Zhang and Bareinboim (2017) and reinforcement learning Buesing et al. (2018); Oberst and Sontag (2019). In the sequel, we detail the SCM for addressing OPPE and explicitly express evaluation of the target policy as an intervention.

In MDP (POMDP) or bandits problems, the casual relationships among variables are straightforward and we only need to care about those unobserved confounders when con-

CFI	OPPE
observations	off-policy episodes
interventions	target policy
query variable	expected return

Table 1: Correspondence between Counterfactual Inference (CFI) and Off-Policy Policy Evaluation (OPPE)

structing SCMs. For example, in POMDP, the agent make a decision based on the history he has observed, which consists of the observation at each time step and the observation is determined by the state the agent is currently in. Meanwhile, the decision taken at this time step will lead the agent to the next state, which in turn influence future's observations. The DAG for this POMDP example is provided in Fig2, where the unobserved confounders in POMDP are also included. According to the DAG, in the SCM, we can express the transition kernel $P(S_t, A_t, S_{t+1})$ as deterministic functions with independent noise variables U, such as $S_{t+1} = f_{st}(S_t, A_t, U_{st})$, which is always possible using auto-regressive uniformization Buesing et al. (2018). Accordingly, we express a given policy π as a causal mechanism: $A_t = f_\pi(H_t, U_{at})$, therefore, Running the target policy π instead of the behavioral policy μ in the environment can be viewed as an intervention $I(\mu \to \pi)$ consisting of replacing $A_t = f_\mu(H_t, U_{at})$ by $A_t = f_\pi(H_t, U_{at})$. We summarize the model-free approach by the following algorithm.

Algorithm 1 Counterfactual Policy Evaluation Buesing et al. (2018)

procedure CFI(observations $\hat{x}_0$, SCM $\mathcal{M}$, intervention I, query variable X_q) ▷ Tis procedure performs counterfactual inference based on the SCM and observations
  $p(u|\hat{x}_0) \leftarrow (\mathcal{M}, \hat{x}_0)$ ▷ Compute the posterior based on the observations and the SCM
  $\hat{u} \sim p(u|\hat{x}_0)$ ▷ Sample noise from the posterior
  $p(u) \leftarrow \delta(u - \hat{u})$ ▷ use the sample distribution as the noise distribution in SCM
  $f \leftarrow f^I$ ▷ perform the intervention and change the causal mechanisms accordingly
  return $x_q \sim p^{\mathrm{do}(I)}(x_q|\hat{u})$
procedure CFI-OPPE(SCM $\mathcal{M}$, target policy π, historical dataset D, number of samples N) ▷ This procedure performs OPPE based on CFI
  for $i \in \{1, 2, \cdots, N\}$ **do**
    $\hat{h}_T^i \sim D$ ▷ sample from the dataset
    $g_i = \mathrm{CFI}\left(\hat{h}_T^i, \mathcal{M}, I(\mu \to \pi), G\right)$ ▷ Counterfactual Inference
  return $\frac{1}{N}\sum_{i=1}^{N} g_i$

4.2.2 Model-based: Counterfactually-guided Model Learning

Different from model-free methods that completely bypass MDP models, model-based ones still relies on the underlying MDP model learned from the observational data generated from the behavioral policy for evaluating the target policy. One of the key challenges in model leaning is the choice of the loss function for model learning without the knowledge of the

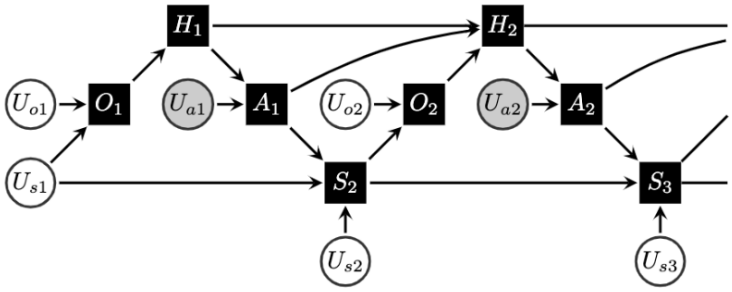


Figure 2: DAG for the POMDP example, taken from Buesing et al. (2018): S_t denotes the state at time t. $H_t := (O_1, A_1, O_2, A_2, \cdots, A_{t-1}, O_t)$ is the set of historical observations up to time t. U_{o_t}, U_{s_t} denote the unobserved confounders for the observation O_t and S_t respectively.

evaluation policy (or the distribution of the evaluation policies). Without this knowledge, we may select a loss function that focuses on learning the areas that are irrelevant for the evaluation policy. Different from previous model-based methods which only take into account data generated by behavioral policy, a new loss function is proposed involving counterfactual thinking when learning the representation using neural networks Liu et al. (2018), which bakes target policy evaluation into representation learning process. This make the learned model more suitable for OPPE, compared with models only involves behavioral data.

To be more specific, for an unknown MDP model $M = \langle r(s,a), P(s,a,s')\rangle$, we can learn a model $\widehat{M} = \left\langle \widehat{r}(s,a), \widehat{P}\left(s,a,s'\right)\right\rangle = \langle h_r(\phi(s),a), h_P\left(\phi(s), a, \phi(s')\right)\rangle$, which is based on the representation function ϕ, i.e., neural networks. Here, $\phi : \mathcal{S} \to \mathcal{Z}$ is a reversible and twice-differentiable function, where $\mathcal{Z}$ is the representation space. The procedure of model-based approach comprises two steps: we first learn the representation ϕ based on the historical data and a carefully selected loss function, derived from counterfactual thinking. Then, we evaluate the target policy using the learned model $\widehat{M}$, which is the same as the classical model-based methods. In this subsection, we mainly focus on the first step, presenting how counterfactual thinking leads to a suitable representation for evaluating the target policy in model learning. Before we move to the details, we first introduce the following notations. For $\tau = (s_0, a_0, \ldots, s_H)$, a trajectory of $H+1$ states and actions from model M, generated by the policy μ, we define the joint distribution of τ as $p_{M,\mu} = d_0\left(s_0\right)\prod_{t=0}^{H-1}\left[P(s_t,a_t,s_{t+1})\mu\left(s_t,a_t\right)\right]$, where d_0, as introduced at the beginning of this section, is the initial state distribution. Accordingly, we can define the associated marginal and conditional distributions as $p_{M,\mu}\left(s_0\right), p_{M,\mu}\left(s_0,a_0\right)$. With these distributions, we can define the $t-$step value function of policy μ as $V^{\mu}_{M,t}(s) = \mathbb{E}_{\tau\sim p_{M,\mu}(s_0)}\{\sum_{i=0}^{t} a_i\}$. Similarly, for the estimated model $\widehat{M}$ and the target policy π, we can define distributions and value functions in the same way.

Now we are in a position for illustrating the combination of counterfactual thinking and mode-based OPPE that leads to a more data-efficient framework for evaluating the target policy, which is first introduced in Liu et al. (2018). Specifically, the key of this framework is that this combination provides a loss function that takes target policy into consideration by counterfactual inference, resulting in a learned model that is suitable for evaluating the tagret

policy. Theoretically, for the behavioral policy μ and the target policy π, the actual difference is measured by $\mathbb{E}_{s_0\sim d_0}\left(V_M^{\pi}\left(s_0\right)-V_M^{\mu}\left(s_0\right)\right)^2$. However, we do not know the underlying MDP model M and can only estimate it based on history data, which we shall detail in the sequel. For now, we assume a learned model $\widehat{M}$ is available, then OPPE relies on the estimated difference$(V_{\widehat{M}}^{\pi}\left(s_0\right)-V_{\widehat{M}}^{\mu}\left(s_0\right))$. One natural question is how different are the actual difference $\left(V_M^{\pi}\left(s_0\right)-V_M^{\mu}\left(s_0\right)\right)$ and the estimated one $(V_{\widehat{M}}^{\pi}\left(s_0\right)-V_{\widehat{M}}^{\mu}\left(s_0\right))$? To answer this question, we first look at the following upper-bound, which is straightforward by using Cauchy-Schwartz inequality.

$$\frac{1}{2}\mathbb{E}_{s_0\sim d_0}\left[(V_{\widehat{M}}^{\pi}(s_0)-V_{\widehat{M}}^{\mu}\left(s_0\right))-(V_M^{\pi}(s_0)-V_M^{\mu}(s_0))\right]^2\leq \mathrm{MSE}_{\pi}+\mathrm{MSE}_{\mu},$$

where $\mathrm{MSE}_{\pi}:=\mathbb{E}_{s_0\sim d_0}\left(V_{\widehat{M}}^{\pi}\left(s_0\right)-V_M^{\pi}\left(s_0\right)\right)^2$ and $\mathrm{MSE}_{\mu}:=\mathbb{E}_{s_0\sim d_0}\left(V_{\widehat{M}}^{\mu}\left(s_0\right)-V_M^{\mu}\left(s_0\right)\right)^2$ are the mean-squared error for two policy value estimates. The above inequality tells that the difference is upper-bounded by the sum of two MSEs, and once we can find a representation ϕ such that this upper-bound $\mathrm{MSE}_{\pi}+\mathrm{MSE}_{\mu}$ is minimized with respect to the representation, then the estimated difference serves as a fairly good evaluation of the target policy. Therefore, the problem reduces to express the upper-bound as a functional of ϕ, since $\widehat{M}$ is merely a notation and it is the reward function $\hat{r}(s,a)=h_r(\phi(s),a)$and the transition kernel $\hat{P}=h_P\left(\phi(s),a,\phi(s'\right)$ that explicitly depend on the ϕ. This is where counterfactual inference plays a part, as we shall see in the following.

For $\mathrm{MSE}_{\mu}=\mathbb{E}_{s_0\sim d_0}\left(V_{\widehat{M}}^{\mu}\left(s_0\right)-V_M^{\mu}\left(s_0\right)\right)^2$, we know the behavioral policy and have the sampled trajectories, hence for this on-policy case, we can apply the Simulation Lemma Kearns and Singh (2002), showing that this on-policy mean square error cane be upper bounded by a function of the reward and transition prediction losses. This upper-bound is provided in the following inequality.

$$\mathbb{E}_{s_0}\left[V_{\widehat{M}}^{\mu}\left(s_0\right)-V_M^{\mu}\left(s_0\right)\right]^2\leq 2H\sum_{t=0}^{H-1}\mathbb{E}_{s_t,a_t\sim p_{M,\mu}}\left[\bar{\ell}_r\left(s_t,a_t,\widehat{M}\right)+\bar{\ell}_T\left(s_t,a_t,\widehat{M}\right)\right],$$

where $\bar{\ell}_r\left(s_t,a_t,\widehat{M}\right)$ and $\bar{\ell}_T\left(s_t,a_t,\widehat{M}\right)$ are the prediction losses of the reward function and the transition kernel respectively and are defined as follows

$$\begin{aligned}
&\bar{\ell}_r\left(s_t,a_t,\widehat{M}\right):=\left(\widehat{r}\left(s_t,a_t\right)-r\left(s_t,a_t\right)\right)^2,\\
&\bar{\ell}_T\left(s_t,a_t,\widehat{M}\right):=\left(\int_{\mathcal{S}}\left(\widehat{P}\left(s_t,a_t,s'\right)-P\left(s_t,a_t,s'\right)\right)V_{\widehat{M},H-t-1}^{\mu}\left(s'\right)ds'\right)^2.
\end{aligned}$$

However, the difficulty here is that we do not have on-policy data for the target policy MSE_{π} and this is why we need counterfactual thinking for constructing an upper bound for MSE_{π}. In Liu et al. (2018), authors decompose the value function error into a one-step reward loss, a transition loss and a next step value loss, with respect to the on-policy distribution. Then, this becomes a contextual bandit problem, and estimation of the MSE is equivalent to the estimation of treatment effect, where the intervention is replacing the behavioral policy μ with the target policy π. This approach is built upon Shalit et al. (2017) about

binary action bandits, where the distribution mismatch is bounded by a representation distance penalty term. Here we merely provide the high-level analysis for estimating MSE_π using causal inference, for more details, we refer readers to Liu et al. (2018); Shalit et al. (2017); Johansson et al. (2016). We consider the following $H-t$ step value error, which decompose the MSE stage-wise, $\epsilon_V(\widehat{M}, H-t) := \int_{\mathcal{S}} \bar{\ell}_V(s_t, H-t)\, p_{M,\mu}(s_t|a_{0:t-1}=\pi)\, ds_t$. In this definition, $\bar{\ell}_V(s, H-t) := \left(V^{\pi}_{\widehat{M},H-t}(s) - V^{\pi}_{M,H-t}(s)\right)^2$ is the squared error of $H-t$ value function and $p_{M,\mu}(s_t|a_{0:t-1}=\pi)$ is the key for constructing the upper-bound for $\epsilon_V(\widehat{M}, H-t)$ recursively and we shall detail it as follows. First of all, $p_{M,\mu}$, as defined above, is the distribution over trajectories under model M and policy μ and $a_{0:t-1}=\pi$ denotes a sequence of actions induced by policy π, i.e., $a_0 = \pi(s_0), \ldots, a_{t-1} = \pi(s_{t-1})$. From this definition, it is easy to see that $\epsilon_V(\widehat{M}, H) = \text{MSE}_\pi$. Denote a factual sequence to be a trajectory that matches the evaluation policy, and let $p^{\phi,F}_{M,\mu}(z_t) = p^{\phi}_{M,\mu}(z_t|a_{0:t}=\pi)$, where F denotes the factual and ϕ means that this distribution is based on the representation and $z_t = \phi(s_t)$. The reason why we consider the representation here is that we aim to construct the upper-bound that involves the representation so that this bound can be used as a loss function for learning the representation. Similarly, we create a counterfactual action sequence with last one action different from $\pi(s_t)$, i.e., $a_{0:t-1}=\pi, a_t \neq \pi$ and accordingly, we let $p^{\phi,CF}_{M,\mu}(z_t) = p^{\phi}_{M,\mu}(z_t|a_t \neq \pi, a_{0:t-1}=\pi)$. If we decompose the whole trajectory into state-action-reward tuples, we can treat this estimation in reinforcement learning as a treatment effect estimation in a sequence of contextual bandits problems as discussed in Shalit et al. (2017) and hence $\epsilon_V(\widehat{M}, H-t)$ can be estimated recursively, as we apply the estimation of each contextual bandit stage-wise to the whole trajectory. The recursive form is given as below.

$$\begin{aligned}\epsilon_V(\widehat{M}, H-t) \leq & 2(H-t)\int_{\mathcal{S}}\left[\bar{\ell}_r\left(s_t, \pi(s_t), \widehat{M}\right) + \bar{\ell}_T\left(s_t, \pi(s_t), \widehat{M}\right)\right] p_{M,\mu}(s_t|a_{0:t}=\pi)\, ds_t \\ & + \frac{H-t}{H-t-1}\epsilon_V(\widehat{M}, H-t-1) + 2(H-t)B_{\phi,t}\,\text{IPM}_{G_t}\left(p^{\phi,F}_{M,\mu}(z_t), p^{\phi,CF}_{M,\mu}(z_t)\right),\end{aligned}$$

where $B_{\phi,t}$ is a constant related to ϕ and t and $\text{IPM}_{G_t}(\cdot,\cdot)$ is the Integral Probability Metric (IPM), measuring the distance of the factual and the counterfactual distribution with respect to the function class G_t. Mathematically, IPM is defined as $\text{IPM}_G(p,q) := \sup_{g\in G}\left|\int g(x)(p(x)-q(x))dx\right|$. Finally, as we estimate $\text{MSE}_\pi = \epsilon_V(\widehat{M}, H)$ using the recursive form, we combine the upper-bounds for MSE_μ and MSE_π and make the combination the loss function for model learning, which leads to a MDP model that is suitable for off-policy evaluation.

4.3 Analysis

If we dive deeper into these causal approaches, we can find that it is the generalization brought up by causal inference that helps solve OPPE problems. Intrinsically, dealing with off-policy evaluation is nothing else but making a prediction about the decision-making system under counterfactual interventions. From this perspective, we argue that causality enhance the generalization of previous off-policy evaluation methods by transferring the cause-effect relationship we infer from observations to the target policy evaluation process. In other words, we are able to handle the "missing data", i.e., the sate-action-reward tuples

included in the target policy that are absent from the sample trajectories generated by the behavioral policy. Specifically, in this subsection, we compare causality-based approaches with existing off-policy evaluation techniques as we briefly introduced at the beginning of this section, showing that how causality equips these novel approaches with better generalization.

Traditional model-free methods, as we have briefly reviewed above, are rather restrictive in the sense that they often need absolute continuity assumption for computing the importance ratio $\rho_t = \pi(s_t, a_t)/\mu(s_t, a_t)$ and besides this, as we can see from the definition of ρ_t, this ratio requires the explicit knowledge of the behavioral policy, i.e., the distribution regarding μ. However, in practice, it so happens that we either do not know the behavioral policy μ or the sample trajectories are in fact generated by multiple behavioral policies. Moreover, absolute continuity is usually not satisfied in real-world applications, which is a major limit of its wider applications, since there is no generalization within this model-free framework. Even though, as mentioned in Jiang and Li (2016); Farajtabar et al. (2018), when the importance ratio is not defined or the behavioral policy is unknown, they can be estimated from the data using some parametric function classes and a good example of estimating the behavioral policy is provided in Farajtabar et al. (2018), we think that these remedies do not address the poor generalization of importance sampling directly, which heavily depends on the quality of observations, since they merely try to create the missing pieces of importance sampling.

Causality make it possible that estimators that do not rest on MDP models are still able to generalize well by incorporating SCMs. Though one may argue that this is not model-free anymore, as this approach involves SCMs, we believe that requiring causal models is the price we have to pay for bypassing the MDP models while acquiring generalization ability, which we shall illustrate later. We first clarify that the term "model-free" indicates that this causal approach circumvents complicated MDP models involving both reward and transition kernel estimation and turns to causal models, which are much easier to deal with, since causal relationships are straightforward in MDP setting, e.g., it is the last sate and action as well as unobserved confounders that lead to the current state. Now we are in a position to elaborate on the advantages of this causality-based model-free approach, as shown below.

- **Generalization** As we have introduced above, this causal approach do not address the distribution mismatch by importance sampling, instead it first provides a SCM, based on which a counterfactual inference is conducted to evaluating the target policy. In other words, as long as the historical observations produce a good estimate of the unobserved confounders and causal mechanisms, we do not care whether absolute continuity is satisfied or not, since it is no longer the key of the estimation. This generalization from observed actions in sample trajectories to counterfactual actions from the target policy is essentially based on the estimated SCM .

- **Unbiased Estimation** Another advantage of this causal approach is it gives unbiased estimates, if the trajectory distribution of the behavioral policy entailed by the SCM coincides with the actual one distribution entailed by the MDP. This shows that causality-based approach returns a consistent evaluation of the target policy, which is different from MDP model-based estimation.

However, a successful implementation of this model-free usually requires the following crucial assumptions. 1) There are no additional hidden confounders, otherwise the current SCM and

CFI are invalid. 2) Causal Mechanisms such as transition kernel and reward functions are easy to model and 3) the inference over the noise U given data to be sufficiently accurate. The first assumption is merely a technical one while the second one can be carefully handled by parametric function classes, e.g., neural nets. The last one plays a significant role, and as discussed in Buesing et al. (2018), imperfect inference over the scenario U could result in wrongly attributing a negative outcome to the agent's actions, instead environment factors. Hence further investigation is needed for measuring the quality of inference.

On the other hand, different from traditional model-free methods, most current model-based methods comprising model learning/estimation process inherently possess generalization to some extent, since the learned model parameters can be directly leveraged when dealing with a new policy without any alterations. However, without causality, we are unable to acquire some prior knowledge about the target policy and generalize from observations, since the common way for designing the loss function is to look at the empirical risk of evaluating the behavioral policy using the model estimation, which reduces the problem to the on-policy model learning. As discussed in Liu et al. (2018), this kind of model learning yields a model that is suitable for evaluating the behavioral policy instead of the target policy, since the representations are chosen in such a way that the estimation error of the behavioral policy is minimized. Therefore, in order to learn a MDP model that is suitable for off-policy evaluation, the target policy must be taken into account when designing the model learning, e.g., the loss function. To be more specific, this causality-based approach measures the mean square error of the target policy by investigating the treatment effect as introduced in the second section. In other words, the difference between one treatment, i.e, the target policy and another treatment, i.e., the behavioral policy is measured by causal inference and is further leveraged to design the loss function, as the treatment effect upper bounds the mean square error. To wrap it up, causality enhance the generalization of model-based approaches in the sense that it helps the model generalize in a desired direction, producing a more accurate off-policy evaluation.

5. Other connections

References

E. Bareinboim, A. Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1342–1350. Curran Associates, Inc., 2015. URL `http://papers.nips.cc/paper/5692-bandits-with-unobserved-confounders-a-causal-approach.pdf`.

R. Bellman. A markovian decision process. *Indiana Univ. Math. J.*, 6:679–684, 1957. ISSN 0022-2518.

L. Buesing, T. Weber, Y. Zwols, S. Racaniere, A. Guez, J.-B. Lespiau, and N. Heess. Woulda, coulda, shoulda: Counterfactually-guided policy search. *arXiv preprint arXiv:1811.06272*, 2018.

D. M. Chickering, D. Heckerman, and C. Meek. Large-sample learning of bayesian networks is np-hard. *Journal of Machine Learning Research*, 5(Oct):1287–1330, 2004.

M. Dudík, J. Langford, and L. Li. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 1097–1104, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195.

O. D. Duncan. *Introduction to structural equation models*. Elsevier, 2014.

M. Farajtabar, Y. Chow, and M. Ghavamzadeh. More robust doubly robust off-policy evaluation. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1447–1456, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL `http://proceedings.mlr.press/v80/farajtabar18a.html`.

S. Fujimoto, E. Conti, M. Ghavamzadeh, and J. Pineau. Benchmarking batch deep reinforcement learning algorithms. *arXiv preprint arXiv:1910.01708*, 2019.

A. S. Goldberger. Structural equation models: An overview. *Structural equation models in the social sciences*, pages 1–18, 1973.

Z. Guo, P. S. Thomas, and E. Brunskill. Using options and covariance testing for long horizon off-policy policy evaluation. In *Advances in Neural Information Processing Systems*, pages 2492–2501, 2017.

M. Jaderberg, W. M. Czarnecki, I. Dunning, L. Marris, G. Lever, A. G. Castaneda, C. Beattie, N. C. Rabinowitz, A. S. Morcos, A. Ruderman, et al. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364(6443): 859–865, 2019.

N. Jiang and L. Li. Doubly robust off-policy value evaluation for reinforcement learning. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 652–661, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `http://proceedings.mlr.press/v48/jiang16.html`.

F. Johansson, U. Shalit, and D. Sontag. Learning representations for counterfactual inference. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 3020–3029, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `http://proceedings.mlr.press/v48/johansson16.html`.

H. Kahn and A. W. Marshall. Methods of reducing sample size in monte carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278, 1953.

M. Kearns and S. Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.

J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

S. Lange, T. Gabel, and M. Riedmiller. Batch reinforcement learning. In *Reinforcement learning*, pages 45–73. Springer, 2012.

L. Li. A perspective on off-policy evaluation in reinforcement learning. *Frontiers of Computer Science*, 13(5):911–912, 2019.

Y. Liu, O. Gottesman, A. Raghu, M. Komorowski, A. A. Faisal, F. Doshi-Velez, and E. Brunskill. Representation balancing mdps for off-policy policy evaluation. In *Advances in Neural Information Processing Systems*, pages 2644–2653, 2018.

S. Mannor, D. Simester, P. Sun, and J. N. Tsitsiklis. Bias and variance approximation in value function estimates. *Management Science*, 53(2):308–322, 2007.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

R. E. Neapolitan et al. *Learning bayesian networks*, volume 38. Pearson Prentice Hall Upper Saddle River, NJ, 2004.

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

C. Paduraru. *Off-policy evaluation in Markov decision processes*. PhD thesis, 2013.

D. Precup, R. S. Sutton, and S. Singh. Eligibility traces for off-policy policy evaluation. In *ICML'00 Proceedings of the Seventeenth International Conference on Machine Learning*, 2000.

D. B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

A. E. Sallab, M. Abdou, E. Perot, and S. Yogamani. Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, 2017(19):70–76, 2017.

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3076–3085. JMLR. org, 2017.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

J. Splawa-Neyman, D. M. Dabrowska, and T. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, pages 465–472, 1990.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

G. Tesauro. Temporal difference learning and td-gammon. *Communications of the ACM*, 38 (3):58–68, 1995.

P. Thomas and E. Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2139–2148, New York, New York, USA, 20–22 Jun 2016. PMLR. URL `http://proceedings.mlr.press/v48/thomasa16.html`.

P. S. Thomas and B. Okal. A notation for markov decision processes. *arXiv preprint arXiv:1512.09075*, 2015.

P. S. Thomas, G. Theocharous, and M. Ghavamzadeh. High-confidence off-policy evaluation. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

Y. Wu, G. Tucker, and O. Nachum. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

J. Zhang and E. Bareinboim. Transfer learning in multi-armed bandit: a causal approach. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*, pages 1778–1780, 2017.