TECHNICAL REPORT
R-65
First version: Dec, 2024
Last version: Mar, 2025

An Introduction to Causal Reinforcement Learning

Elias Bareinboim EB@CS.COLUMBIA.EDU
Junzhe Zhang JUNZHEZ@CS.COLUMBIA.EDU
Department of Computer Science
Columbia University
New York, USA
Sanghack Lee SANGHACK@SNU.AC.KR
Graduate School of Data Science
Seoul National University
Seoul, Republic of Korea

Abstract

Causal inference provides a set of principles and tools that allows one to combine data and knowledge about an environment to reason with questions of counterfactual nature – i.e., what would have happened had reality been different – even when no data of this unrealized reality is currently available. Reinforcement learning provides a collection of methods to learn a policy that optimizes a specific measure (e.g., reward, regret) when the agent is deployed in an environment and pursues an exploratory, trial-and-error approach. These two disciplines have evolved independently and with virtually no interaction between them. We note that they operate over different aspects of the same building block, i.e., counterfactual relations, which makes them umbilically connected. Based on these observations, we further realize that various novel learning opportunities naturally arise when this connection is explicitly acknowledged, understood, and mathematized. To realize this potential, we further note that any environment where the RL agent is deployed can be decomposed as a collection of autonomous mechanisms that lead to different causal invariances and which can be parsimoniously modeled as a structural causal model; any standard RL setting today is implicitly encoding one of these models. This natural formalization, in turn, will allow us to put under a unifying treatment different modes of learning, including online, off-policy, and causal calculus learning, which appear seemingly unrelated in the literature. One may surmise that these three standard learning modalities are exhaustive in the sense that all possible counterfactual relations are learnable through their continuous implementation. We show that this is not the case by introducing several quite natural and pervasive classes of learning settings that do not fit these modalities but entail novel dimensions and types of analysis. Specifically, we will introduce and discuss through causal lenses the problem of generalized policy learning, where to intervene, imitation learning, and counterfactual learning. This new set of tasks and understanding lead to a broader view of counterfactual learning and suggests the great potential for the study of causal inference and reinforcement learning side by side, which we call *causal reinforcement learning* (CRL).

Keywords: Structural Causal Models, Interventions, Counterfactuals, Reinforcement Learning, Identifiability, Robustness, Generalizability, Off-policy Evaluation, Imitation Learning.

1. Introduction

AI will play an increasingly prominent role in society as significant portions of its decision-making infrastructure are being delegated to automated systems. The transition from human-based to AI-

based decision-making systems is underway and will likely accelerate in the coming years. The new generation of AI systems is expected to be more efficient, robust, explainable, generalizable, and, more importantly, to lead to outcomes aligned with society's goals and expectations. There is a growing understanding that robust decision-making relies on some knowledge of the environment's underlying causal mechanisms. For instance, an intelligent robot needs to know the cause-and-effect relationships in its environment to plan its course of actions more robustly and communicate them with humans; a physician needs to understand the individual and joint effects of multiple available drugs to design an effective strategy for her patients while avoiding unethical experimentation with human subjects and potentially harmful side effects; an economist needs to understand the relationship between skill sets and the future job market demands so as to design new training and educational policies more efficiently. These examples of everyday decision-making found across society rely on some understanding of the often complex, dynamic, and almost invariably unobserved collection of causal mechanisms.

One of the primary goals and unifying themes found across Artificial Intelligence (AI) is to develop a *rational agent* that operates in an *environment* capable of maximizing a performance measure, and based on the prior knowledge that the agent has about it (Russell, 2010; Sutton and Barto, 1998). Operationally, an agent perceives the environment's state through sensory input (e.g., cameras, lidar, API) and then interacts with it through (physical or virtual) *actuators*. The agent usually follows what is known as a *policy*, a sequence of decision rules that dictates the action based on the evolving history of its perception and prior decisions. When the underlying system dynamics are provided *a priori* (e.g., in the form of parameters set), one could obtain an optimal policy by applying standard planning algorithms (Bellman, 1957; Puterman, 1994; Sutton and Barto, 1998; Bertsekas, 2005). Effective planning methods in structured environments are studied under the rubrics of influence diagrams (Shachter, 1986; Lauritzen and Nilsson, 2001; Koller and Milch, 2003). In many practical applications, however, the parameters of the real, underlying environment are not fully known, which entails some learning processes.

Reinforcement learning (RL) has become the *de facto* framework for reasoning about optimal decision-making under uncertainty in AI and machine learning over the last decades. RL methods could generally be categorized according to the types of interactions invoked during the learning stage between the agent and the environment. First, there is the modality of *off-policy learning* where the agent learns an optimal policy from offline data generated by a different behavior policy or agent. Second, there is *online learning*, where the agent directly deploys policies in the actual environment and observes subsequent outcomes. Several RL algorithms have been proposed on the formalism (and the corresponding assumptions) known as Markov Decision Processes (MDPs), where a finite set of state variables is statistically sufficient to summarize the treatments and covariates history (Bellman, 1957; Puterman, 1994; Sutton and Barto, 1998; Bertsekas, 2005; Jaksch et al., 2010). There are a number of variations of this setting – both special cases and generalizations – including multi-armed bandits (MABs) (Thompson, 1933; Robbins, 1952; Lai and Robbins, 1985b; Auer et al., 2002a), contextual banduits (Auer et al., 2002c; Langford and Zhang, 2008b), and partially-observable MDP (POMDPs) (Lovejoy, 1991; Singh et al., 1994; Jaakkola et al., 1995; Littman et al., 1995),

As noted by Bellman, the "curse of dimensionality" is a pervasive challenge found in the RL literature, considering the exponential growth in the state-action space when solving the corresponding dynamic programming problems (Bellman, 1957). Since real-world problems often involve more than a few variables, it would appear unlikely that RL methods could be applied to practical settings

of higher significance. Fortunately, the real world is often structured and modular, and components are usually independent of each other in most systems. Hierarchical RL leverages the independence relationships and allows the decomposition of the MDP model into a hierarchy of simpler MDPs representing subtasks (Dietterich, 2000; Guestrin et al., 2001, 2003; Kulkarni et al., 2016). Consequently, by implicitly exploiting the modularity following from the causal mechanisms, the computational complexity of solving RL tasks could be significantly reduced. More recently, the use of deep learning automates the learning of lower-dimensional and hierarchical representations in RL, including deep Q-networks (Mnih et al., 2015), AlphaGo (Silver et al., 2016), and IBM's Watson DeepQA system (Ferrucci et al., 2010).

Given the compelling results obtained so far, one may surmise that the foundational picture is essentially complete. In other words, all decision-making problems would eventually be solved given additional resources – more data, more memory, and more computation. We will argue in this part of the book that this is not the case. In fact, researchers are becoming increasingly aware that AI systems deployed in the real world are very fragile, brittle, and are commonly sample inefficient ("data-hungry"). They also lack robustness and generalizability capabilities, and are opaque, being neither interpretable nor explainable, and subsequently, not trustworthy from a human perspective. This observation does not constitute a fundamental impediment to their evolution, of course, but is a simple realization of the current state of affairs, which is a necessary first step toward finding a solution. As it will become apparent throughout our exposition, underlying these challenging issues is the lack of an explicit language capable of accounting for the causal mechanisms and exogenous sources of variations that amount to the environment where the agent is deployed, and fundamentally generate the invariances, rewards, and dynamics of the system.

Interestingly, the field of causal inference (CI) provides a set of principles and tools that allows one to combine data and structural invariances about the environment to reason about questions of counterfactual nature, i.e., what would have happened had reality been different, even when no data about this imagined reality is available, and the environment is not fully observable. The conditions under which the effect of an action can be computed from observational and experimental data have been extensively studied. Several conditions and algorithms have been proposed based on qualitative knowledge about the environment (Part II). Notable results exist throughout the empirical sciences (Cornfield, 1951; Flegal et al., 2005; Heckman, 2006; U.S. Department of Health and Human Services, 2014), and more recently, in machine learning (Kallus and Zhou, 2018; Namkoong et al., 2020; Etesami and Geiger, 2020; Kallus and Zhou, 2020; Tennenholtz et al., 2021) on how to translate causal knowledge to support new policies and principled decision-making. However, this area is still in its infancy, requiring substantive work and a significant amount of highly skilled scientists.[1] Our goal in AI is to create intelligent systems capable of reasoning and autonomous actions. This requires a transition from a heuristic grasp of the interplay between causal knowledge

1. For instance, economists strive to understand the *root causes* of poverty, which could allow, in principle, the design of new policies (i.e., causal interventions) to improve the population's socioeconomic status (SES). A considerable body of evidence was accumulated for many decades, notably by the University of Chicago's Professor and Nobel Prize laureate, James Heckman, who demonstrated the effects of early childhood education on families' SES, among other indicators (Heckman, 2006). The understanding following this causal link translated to the larger support of early childhood education and a push for new policies aligned with these findings; for example, see Obama's one billion dollar investment (The White House, Office of the Press Secretary, 2014). There are many such cases throughout the empirical sciences — e.g., evidence supports that tobacco smoking is one of the determinant factors of lung cancer (Cornfield, 1951; U.S. Department of Health and Human Services, 2014), or obesity is responsible to shortening life expectancy (Flegal et al., 2005), which, in turn, translated into new public health policies.

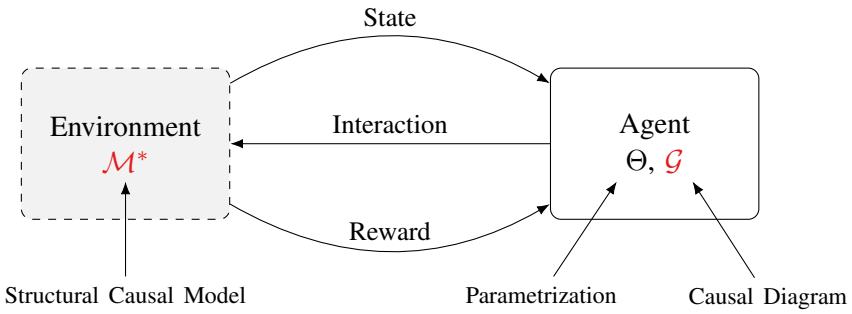


Figure 1: The Agent-Environment interaction from Causal Reinforcement Learning (CRL).

and decision-making to a deeper and more fundamental understanding of the principles that connect causal evidence and robust decision-making under uncertainty.

In this paper, we endeavor towards this aim, harnessing the strengths and synergies of CI and RL to devise more sample-efficient, transparent, and robust decision-making systems. We refer to this program as *Causal Reinforcement Learning* (CRL). Specifically, this chapter seeks to introduce this framework and investigate the intricate and sometimes nuanced relationship between causal knowledge and decision-making. The design of CRL agents will follow two simple and powerful observations. First, the environment's underlying causal mechanisms should be accounted for explicitly in the analysis. In particular, this is realized through formal language and a dual view, as depicted in Fig. 1:

1. From the environment's perspective, causal mechanisms and the probability distribution over the exogenous conditions are described as a fully specified SCM $\mathcal{M}^*$ (on the figure's right side).

2. From the agent's perspective, a parsimonious representation of the environment's invariances will be maintained in the form of a causal model $\mathcal{G}$ (on the left side), such as a causal diagram.

In essence, while the environment and agent's perspectives differ, they are tied through the pair, an SCM $\mathcal{M}^*$ and its corresponding causal diagram $\mathcal{G}$. There is formally a notion of compatibility between these two objects, as discussed later in the book. This pairing can be articulated in different forms, including typical template-like structures such as MABs, MDPs, POMDPs.[2]

The second observation that ground the CRL framework follows from the understanding that every SCM M^* begets a mathematical construct named the *Pearl Causal Hierarchy* (PCH) (Pearl and Mackenzie, 2018), which has been named in his honor and formalized in (Bareinboim et al., 2020). The PCH consists of three qualitatively different types of distributions that are separated into layers – the associational, the interventional, and the counterfactual. The PCH will play a central role in formalizing the types of activities an agent can engage in when considering the environment it has been deployed into, including *seeing*, *doing*, and *imagining*. As illustrated in Table 1, knowledge at each layer will allow the agent to reason about different classes of *causal concepts*, or "queries." In particular, Layer 1 deals with purely "observational", factual information when the agent passively observes the environment (or other agents interacting in the environment). Layer 2 encodes information about what *would* happen, hypothetically speaking, were some interventions

2. The current literature evokes this relationship mostly in an implicit fashion, assuming that there exists a match between the environment and the agent knowledge.

	Layer (Symbol)	Typical Activity	Quintessential Question	Example	Machine Learning
$\mathscr{L}_1$	Associational $P(y\|x)$	Seeing	What is? How would seeing X change belief in Y?	What does an observation tell us about the underlying state?	Supervised / Unsupervised Learning
$\mathscr{L}_2$	Interventional $P(y\|\text{do}(x))$	Doing	What if? What if I do X?	What if I brake hard, will my vehicle avoid an accident?	Reinforcement Learning
$\mathscr{L}_3$	Counterfactual $P(y_x\|x', y')$	Imagining	Why? What if I had acted differently?	Was it the hard brake that prevented the accident?	Explanation Transparency

Table 1: Summary of Pearl's Causal Hierarchy including each of its layers, the symbolic representation, typical activities and questions, and examples where it appears in ML settings.

were performed, namely, the effects of actions. Interestingly, this is a typical activity in RL settings, and answering such queries may be possible from data on interventions already performed or from data collected passively under the first modality in layer 1. These will appear in the form of online and offline learning modalities, discussed later in the text. Finally, Layer 3 involves queries about what *would have* happened, counterfactually speaking, had some interventions or actions been performed, given that something else in fact occurred, possibly conflicting with a hypothetical intervention that has not actually happened. The causal hierarchy establishes a useful classification of concepts that might be relevant for a given CRL inference task, thereby also classifying formal frameworks in terms of the questions that agents are able to represent, and ideally answer.

There is a growing literature that investigates various points in the design space of CRL agents and their policies and represents the more concrete examples currently available of this picture, and CRL tasks and inferential machinery (Bareinboim et al., 2015; Forney et al., 2017; Lee and Bareinboim, 2018b; Kallus and Zhou, 2018; Forney and Bareinboim, 2019; Lee and Bareinboim, 2019b; de Haan et al., 2019; Lee and Bareinboim, 2020; Namkoong et al., 2020; Etesami and Geiger, 2020; Kallus and Zhou, 2020; Bennett et al., 2021; Wang et al., 2021; Tennenholtz et al., 2021; Kumor et al., 2021; Ruan and Di, 2022; Swamy et al., 2022). Still, the treatment provided in each paper represent special cases of specific problems, and were not studied in generality and in a unified manner. In fact, these problems can be seen as a basis on which a CRL agent should be built and will be studied under the same formal umbrella that motivated their very existence.

1.1 Roadmap of the Paper

The remainder of the paper is organized as shown in Fig. 2. We provide in Sec. 2 the necessary background and a logical foundation of causal inference to understand the rest of this paper. We review the definition of structural causal models (Sec. 2.1), evaluation of observational and interventional distributions (Sec. 2.2), and the construction of causal diagrams representing qualitative knowledge in SCMs (Sec. 2.3). Extensive examples are provided to illustrate these concepts.

Sec. 3 is a foundational chapter that connects the different learning modalities found in RL to the causal language introduced in this chapter. In particular, Sec. 3 formalizes the policy learning problem using the semantic language of SCMs, termed causal decision models (Sec. 3.1). Based

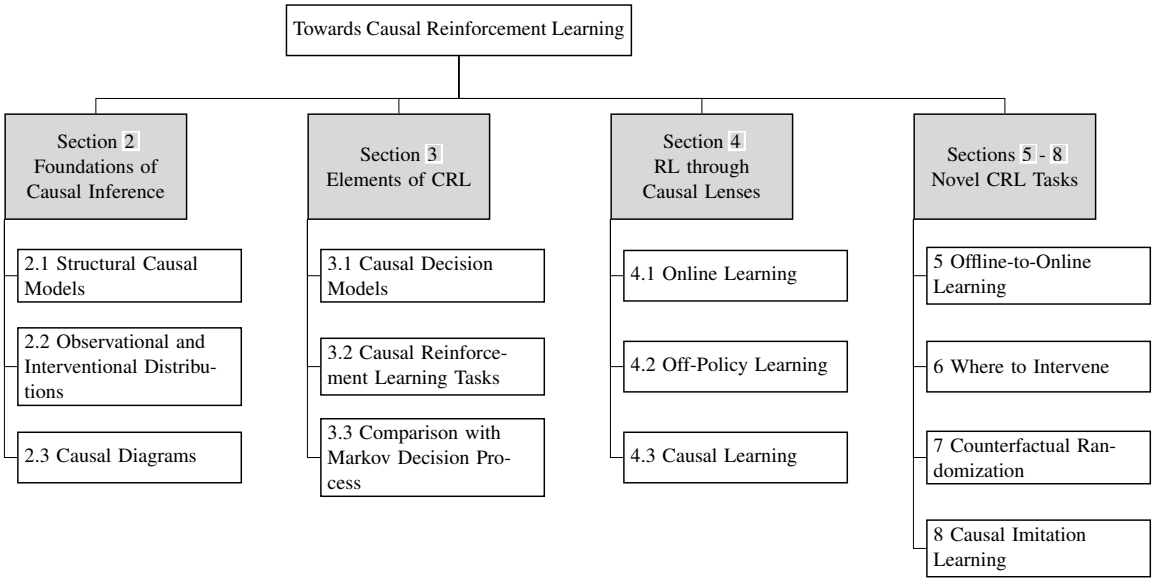


Figure 2: Paper's roadmap and organization.

on this framework, we introduce *causal reinforcement learning tasks* that consider the interaction capabilities of the learning agent and the prior knowledge of the environment accessible to the agent (Sec. 3.2). We compare the CRL formalisms with reinforcement learning under the standard model assumptions of Markov decision processes, emphasizing that there exists no discretion here, and causal knowledge is indispensable for solving CRL tasks.

Sec. 4 studies classic learning tasks of reinforcement learning and causal inference through the CRL framework, including off-policy learning (Sec. 4.1), online learning (Sec. 4.2), and causal identification (Sec. 4.3). In particular, we discuss several conditions and algorithmic procedures for policy learning for each of these tasks. In the last section, we introduce a graphical criterion that extends off-policy learning methods to the language of structural causality where unobserved confounding is not ruled out a priori.

Sec. 5 considers the problem of causal offline-to-online learning (COOL), where the agent attempts to first pre-train informative representations of optimal policies from offline data and then fine-tune policy estimates by conducting online experimentations. Sec. 5.1 introduces a confounding robust procedure for transferring observational data in bandit models. Secs. 5.2 and 5.3 extend this transfer strategy to sequential decision-making settings where the agent has to determine a series of actions in order to maximize the primary outcome (e.g., dynamic treatment regimes).

Sec. 6 introduces a new task called *mixed policy learning*. This task is concerned with whether the agent should intervene in the system, and, if so, where the intervention should be targeted. Sec. 6.1 investigates the structural properties inherent in a mixed policy space with atomic intervention, where the properties can help the agent to explore the space more efficiently and effectively. Sec. 6.2 further investigates a scenario where the agent can conduct soft intervention, selecting which variables to observe for performing soft intervention.

Sec. 7 broadens the scope of policies and introduces a novel *counterfactual decision criterion* that is applicable when agents have their own biases and operate in adversarial environments. Sec. 7.1 formalizes the concept of counterfactual policies, enabling agents to perform counterfactual

reasoning by accounting for their initial intended actions. Sec. 7.2 presents a new type of counterfactual randomization strategy that supports the realization of the counterfactual decision criterion and facilitates the learning of an optimal counterfactual policy. In the last section, we formalize the trade-off between optimality and autonomy under the counterfactual decision criterion and provide a practical planning algorithm to address this trade-off.

Sec. 8 studies the problem of policy learning from the observational data without complete knowledge about the reward function measuring the performance of the agent - called imitation learning. Sec. 8.1 develops a complete graphical condition for learning an imitating policy, achieving the expert's performance using behavioral cloning. Sec. 8.2 extends this condition to produce a policy that could consistently dominate the expert by exploiting parametric knowledge about the unknown reward function via inverse RL. We also develop an algorithmic approach to apply inverse RL in a more generalized family of SCMs provided with a causal diagram of the environment.

Finally, Sec. 9 concludes by summarizing the work and algorithms studied in previous sections and giving final remarks. We also discuss other essential CRL tasks, including transportability, generalizability and model induction, and outlines future challenges in designing CRL agents.

1.2 Notations

We introduce the basic notations used throughout this paper. Capital letters represent variables (X), and small letters represent their values (x). Let $\mathscr{D}(X)$ represent the domain of X and $\mathscr{P}_X$ the space of probability distributions over $\mathscr{D}(X)$. Boldfaced capital letters $\boldsymbol{X}$ denote a collection of variables, $|\boldsymbol{X}|$ its dimension, $\mathscr{D}(\boldsymbol{X})$ their joint domains, and boldfaced smaller letters $\boldsymbol{x}$ a particular joint realization in the domain $\mathscr{D}(\boldsymbol{X})$. We will consistently use $P(\boldsymbol{X})$ to represent the joint distribution over $\boldsymbol{X}$ and $P(\boldsymbol{x})$ represent probabilities $P(\boldsymbol{X}=\boldsymbol{x})$; similarly, notation $P(\boldsymbol{Y} \mid \boldsymbol{X})$ represents a set of conditional distributions $P(\boldsymbol{Y} \mid \boldsymbol{X}=\boldsymbol{x}), \forall \boldsymbol{x}$. Finally, indicator function $\mathbb{1}\{\boldsymbol{Z}=\boldsymbol{z}\}$ returns 1 if $\boldsymbol{Z}=\boldsymbol{z}$ holds true; otherwise $\mathbb{1}\{\boldsymbol{Z}=\boldsymbol{z}\}=0$.

For a directed acyclic graph (DAG) $\mathcal{G}$, we denote by $\boldsymbol{V}(\mathcal{G})$ the set of vertices in $\mathcal{G}$; similarly, $\boldsymbol{E}(\mathcal{G})$ is the set of arrows in $\mathcal{G}$. A vertex-induced subgraph is denoted by brackets, e.g., $\mathcal{G}[\boldsymbol{W}]$ which includes vertices $\boldsymbol{W} \subseteq \boldsymbol{V}(\mathcal{G})$ and edges among its elements. For convenience, we define $\mathcal{G} \setminus \boldsymbol{X} \equiv \mathcal{G}\left[\boldsymbol{V}(\mathcal{G}) \setminus \boldsymbol{X}\right]$. For arbitrary subsets $\boldsymbol{W}, \boldsymbol{Z} \subseteq \boldsymbol{V}(\mathcal{G})$, $\mathcal{G}_{\overline{\boldsymbol{W}}\underline{\boldsymbol{Z}}}$ is a subgraph obtained from $\mathcal{G}$ by removing edges pointing into any node $W \in \boldsymbol{W}$ and edges coming out of any node $Z \in \boldsymbol{Z}$.

Also, we will use standard graph-theoretic abbreviations to represent relationships among nodes: $an(X)_{\mathcal{G}}$, $de(X)_{\mathcal{G}}$, $pa(X)_{\mathcal{G}}$ and $ch(X)_{\mathcal{G}}$ stand for the set of ancestors, descendants, parents, and children of a node X in a DAG $\mathcal{G}$, not including X; subscript $\mathcal{G}$ is omitted when it is obvious. The parent set of a node set $\boldsymbol{X}$ is all parents of any node in $\boldsymbol{X}$, i.e., $pa(\boldsymbol{X})_{\mathcal{G}} = \bigcup_{X \in \boldsymbol{X}} pa(X)_{\mathcal{G}}$; $de(\boldsymbol{X})_G$, $an(\boldsymbol{X})_G$, and $ch(\boldsymbol{X})_G$ are similarly defined. Finally, $Pa(\boldsymbol{X})_{\mathcal{G}}$ is the set union $pa(\boldsymbol{X})_{\mathcal{G}} \cup \boldsymbol{X}$, and so do $De(\boldsymbol{X})_{\mathcal{G}}$, $An(\boldsymbol{X})_{\mathcal{G}}$, and $Ch(\boldsymbol{X})_{\mathcal{G}}$.

2. Foundations of Causal Inference

We now provide a brief outline of this section and will lay out the basic CRL building blocks, following the schema shown in Fig. 3. We will start in Sec. 2.1 with the underlying environment, which will be described through formal causal semantics and the definition of SCMs. We will further provide some specific examples, or instantiations of some SCMs that follow from canonical examples found in the literature (e.g., MABs, MDPs). Each SCM induces the PCH, shown in the bottom of the figure, modeling various interactions an agent can undertake in the environment,

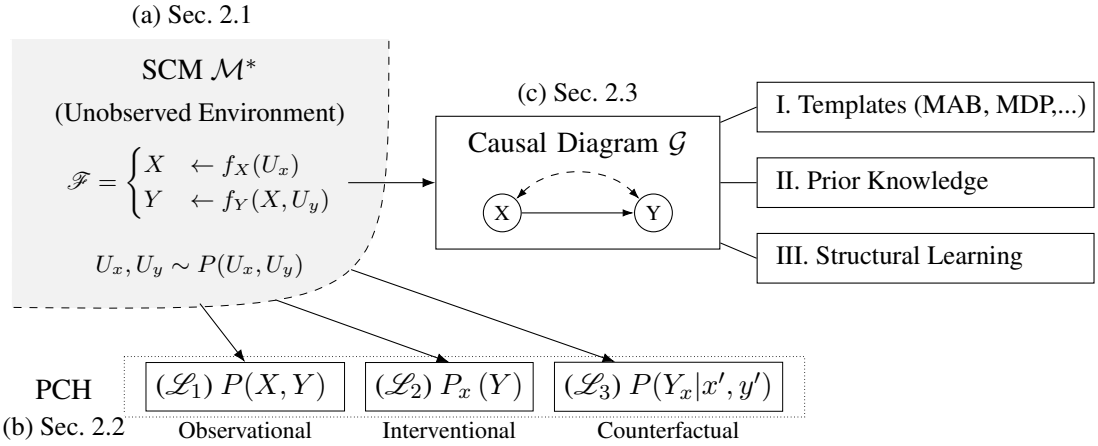


Figure 3: Building blocks of Causal RL analysis. (a) Unobserved model of the environment; (b) the PCH (and not necessarily fully observed); (c) The structural constraints over the SCMs that may be elicited through different methods, including prior knowledge and structural learning.

including observations, interventions, and counterfactual reasoning. In Sec. 2.2, we will formalize the dynamics when the agent is only passively observing the environment – i.e., collecting the PCH's $\mathscr{L}_1$-type of data – which will give rise to what is known as the *observational distribution*. We will then move to a more active mode of interaction whenever the agent can perform interventions in the environment, which constitute $\mathscr{L}_2$-type of interactions, giving rise to what is known as the *interventional distribution*. In Sec. 2.3, we discuss specifying structural assumptions about the environment in a non-parametric and parsimonious manner using causal diagrams.

2.1 Structural Causal Models and the Environment

We build on the language of *structural causal models*, which is one of the most general and flexible data-generating models known to date (Pearl, 2000; Bareinboim et al., 2020). The first element of any CRL system is the environment where the agent will be deployed, which will be instantiated as an SCM defined next.

Definition 1 (Structural Causal Model (Pearl, 2000)) *A structural causal model (SCM, for short) $\mathcal{M}$ is a 4-tuple $\langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P \rangle$ where:*

- *A set of background or exogenous variables $\boldsymbol{U} = \{U_1, U_2, ..., U_k\}$, representing factors outside the model, which nevertheless, affect relationships within the model;*
- *A set of endogenous variables $\boldsymbol{V} = \{V_1, V_2, ..., V_n\}$ representing variables inside the model.*
- *A set $\mathscr{F}$ of structural functions $\{f_i : V_i \in \boldsymbol{V}\}$ s.t. each f_i determines the value of $V_i \in \boldsymbol{V}$,*

$$v_i \leftarrow f_i(\boldsymbol{pa}_i, \boldsymbol{u}_i), \tag{1}$$

  where $\boldsymbol{PA}_i \subseteq \boldsymbol{V} \setminus \{V\}$ and $\boldsymbol{U}_i \subseteq \boldsymbol{U}$.

- *A probability distribution over the exogenous variables, $P(\boldsymbol{U})$.* ■

Some observations follow from this definition. First, note that a typical SCM $\mathcal{M}$ partitions the variables into two sets – exogenous (unobserved) $\boldsymbol{U}$ and endogenous (observed) $\boldsymbol{V}$. The values of exogenous variables $\boldsymbol{U}$ are decided outside the environment where the agent is deployed, following a *probability distribution* $P(\boldsymbol{U})$ over all possible configurations of its states. In practical settings, these variables represent unaccounted factors of the units and of the situations that affect the environment under consideration; for instance, possibly the patients' DNA, customers' sensitive demographics, robot's physical constraints, and other features and states that affect the system but are unobserved from the perspective of the CRL agent. For concreteness, consider a physician who may not have access to the patient's DNA before prescribing a certain treatment. At the same time, such an attribute influences how well the patient will respond to the drug and whether they will recover from their condition. Naturally, other physicians who may have access to this information will use it when making their decisions. Alternatively, a robot may not be able to determine its precise location in the building, while others with more accurate sensors may have an almost perfect reading of their positions. Of course, there is no "right" view of reality, and those are alternative perspectives based on the contingencies and capabilities of each agent and the environment where it is deployed.

More generally, the exogenous set $\boldsymbol{U}$ allows for the existence of *unobserved confounding*, which is inherent in any real-world, complex setting in which not every bit of information can be measured or assessed by the agent. This is a common phenomenon in environments where humans are present since their existence precludes full observability.[3] Unless otherwise specified, our referential frame would always be from the CRL agent's perspective, which will come with their specific perceptual and interventional ($\mathscr{L}_2$) capabilities.

Second, the value of each endogenous variable $V_i \in \boldsymbol{V}$ in $\mathcal{M}$ is determined by a causal mechanism f_i, which takes other endogenous and exogenous variables in the system as input. The randomness of exogenous variables $\boldsymbol{U}$ induces variations in the endogenous variables $\boldsymbol{V}$, which is formalized in the next section. These causal mechanisms, together with the background factors encoded in the distribution $P(\boldsymbol{U})$, represent the data-generating process according to which the environment decides observed states and rewards in each possible configuration.

One feature of the representation of SCMs is its flexibility: it could contain an arbitrary collection of causal mechanisms. This naturally includes and allows one to model any standard RL environment using the SCM framework, including MABs and MDPs, as illustrated below. One important distinction is that standard RL models leave action variables $\boldsymbol{X}$ in the environment uninstantiated. On the contrary, SCMs explicitly model some secondary controlled process associated with actions, such as a human operator or a different source agent demonstrating the task, or even the effect of nature itself (e.g., gravity). Formally, we denote by $f_{\boldsymbol{X}} = \{f_X \mid \forall X \in \boldsymbol{X}\}$ the collection of functional relationships in the system determining values of $\boldsymbol{X}$. By default, the system is said to be under a *behavior policy*, in RL terminology (Sutton and Barto, 1998), or under the natural regime, in CI language (Pearl, 1995, 2000; Dawid, 2002). For example, a physician makes decisions about the patient's treatment in the natural world, regardless of what the AI agent wants to do. Also, the position and momentum of a particle may change with gravity, even though no

3. Some research from Neuroscience suggests that decision-making may be a process handled largely by subconscious mental activity (Libet et al., 1993). Even several seconds before we consciously decide, its outcome can be largely influenced by subconscious activity in the brain. In other words, this suggests that humans generally have a lack of understanding of our decision-making process and have a hard time measuring all factors influencing our behaviors; therefore, full observability is rarely realizable in practice.

deliberate (or human) agent interferes in the system at each instant in time. The main observation relevant to our context is that, regardless of how $\boldsymbol{X}$ attains its value, this does not happen naturally under the control of the CRL agent, which is the one we care about here.

To start understanding Def. 1, we consider some examples showing how some classic, increasingly refined RL environments should be modeled through the semantics of SCMs. We will take a natural regime perspective, which considers the CRL agent (learner) passively observing a different agent (e.g., a teacher, a physical law) making decisions in the environment; these other agents have a possibly different set of perceptual and interventional capabilities.[4]

Example 1 (Multi-Armed Bandit (Robbins, 1952)) *In a clinic where patients with a chronic disease are treated, physicians must choose how to select their treatment, one at a time. There are two treatments or 'arms', $X = 0$ and $X = 1$, and the overall goal is to find the optimal treatment to maximize the chance of patients' recovery (Y). This can be seen as an example of a multi-armed bandit (MAB) model. Its roots can be traced back to work produced by Thompson (1933), which was further developed in (Robbins, 1952; Gittins, 1979; Lai and Robbins, 1985a; Auer et al., 2002a). Consider a MAB model described by an SCM*

$$\mathcal{M}^*_{\text{MAB}} = \langle \boldsymbol{U} = \{U\}, \boldsymbol{V} = \{X, Y\}, \mathscr{F}, P(U) \rangle, \tag{2}$$

where U represents the patient's age (e.g., normalized in a real interval $[0, 1]$). The causal mechanisms are the following:

$$\mathscr{F} = \begin{cases} X \leftarrow \mathbb{1}\{U < 0.8\}, \\ Y \leftarrow \mathbb{1}\{U < 0.4 - \Delta X\} \end{cases} \tag{3}$$

where coefficient Δ is a real number bounded in $(0, 0.4)$; and $P(U)$ is such that values of U are drawn from a uniform distribution $\texttt{Unif}(0, 1)$.

In words, the physician prescribes treatment $X = 0$ for senior patients ($U \geq 0.8$); otherwise, an alternative treatment, $X = 1$, is prescribed. Meanwhile, the recovery Y of the disease depends on the treatment (X) and the patient's age (U). Since the coefficient $\Delta > 0$, the chance of recovery when prescribing $X = 0$ is higher than when prescribing $X = 1$. That is, $X = 0$ is the preferable treatment in this context. ■

Interestingly, note that the SCM given by Eq. 3 contains two qualitatively different types of mechanisms – f_X is controlled by the physician, and f_Y is controlled by Nature, representing in this case, the patient's biology. From the CRL's agent point of view, both mechanisms are external and then deemed as the environment.

Example 2 (Markov Decision Process (Puterman, 1994)) Markov Decision Process *(MDP) has emerged as the* de facto *framework for reasoning about the sequential decision-making in AI (Sutton and Barto, 1998; Russell and Norvig, 2016). An example of MDP is the day-to-day management of an inventory with a fixed maximum size (Szepesvári, 2010).*

4. Since our goal here is not to discuss the subtleties of multi-agent systems, we will abstract away the identity of these other agents and simply use the notion of an environment.

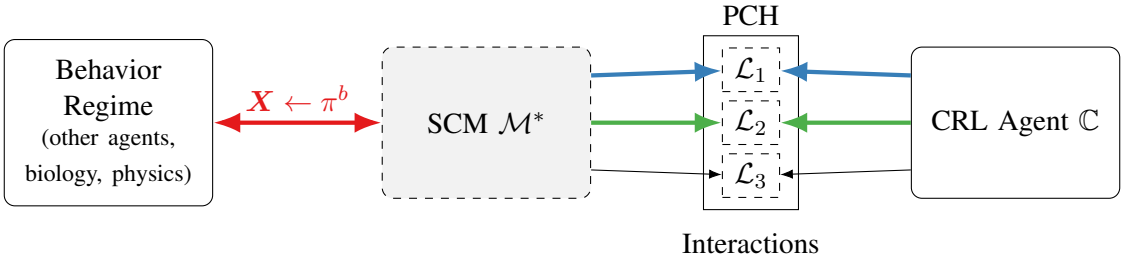


Figure 4: Representation of the CRL agent (right side) interacting with the SCM (middle) through natural (marked in blue) and interventional (green) regimes. Other behavior agents and their interactions are shown in the left side (red).

On day $i = 1, 2, \ldots$, the inventory manager observes the current size of the inventory S_i, decides whether to purchase new items to fill up the inventory X_i, and receives a subsequent profit Y_i by the end of day i. More specifically, consider an MDP model described by the SCM:

$$\mathcal{M}^*_{\text{MDP}} = \langle \boldsymbol{U} = \{U_{i,1}, U_{i,2}, U_{i,3}\}, \boldsymbol{V} = \{X_i, Y_i, S_i\}, \mathscr{F} = \{\mathscr{F}_i\}, P(\boldsymbol{U})\rangle_{i=1,2,\ldots}, \tag{4}$$

where $U_{i,1}, U_{i,2}, U_{i,3}$ represent, respectively, human errors when stocking the inventory, and uncertainties in demand, and monetary values of the goods. The causal mechanisms $\mathscr{F}_i$ representing the system dynamics transitioning from day $i-1$ to i are defined as:

$$\mathscr{F}_i = \begin{cases} S_i \leftarrow (S_{i-1} \vee X_{i-1}) \oplus U_{i-1,1} \oplus U_{i-1,2}, \\ X_i \leftarrow S_i \oplus U_{i,1} \\ Y_i \leftarrow S_i \oplus X_i \oplus U_{i,1} \oplus U_{i,3} \end{cases} \tag{5}$$

and $P(\boldsymbol{U})$ is such that $U_{i,1}, U_{i,2}, U_{i,3} \in \{0, 1\}$ are independent variables drawn from distributions $P(U_{i,1} = 1) = P(U_{i,2} = 0) = P(U_{i,3} = 0) = 0.9$.

In words, the size of current inventory $S_i = 1$ is full if it is also full $S_{i-1} = 1$ on the previous day or gets refilled $X_{i-1} = 1$; the operation error $U_{i-1,1}$ and customers' demand $U_{i-1,2}$ could also affect the inventory size. The manager's decision X_i depend on the inventory size S_i on the day and potential operation errors $U_{i,1}$. The store only makes a profit $Y_i = 1$ if the inventory is well managed: it is refilled when empty, or no new goods is purchased when full. Similarly, operation errors and price fluctuation could also affect the net profit on day i. ■

A couple of observations follow. First, SCMs provide a flexible language that allows for the natural encoding of standard RL environments. Second, the specific instantiation of each SCM will, in general, not be visible to the agent, and the goal is just to represent the environment's generative processes and its implied PCH. Third, the SCM is different than the PCH's corresponding datasets (Sec. 2.2) and the assumptions the agent may make about them (Sec 2.3), as discussed next.

2.2 Learning through Observational & Interventional Regimes (PCH's Layers 1 and 2)

One important feature of the CRL architecture described so far is the explicit division between the agent and the environment where it lives, which induces a natural line between the set of endogenous ($\boldsymbol{V}$) and exogenous ($\boldsymbol{U}$) variables. Interestingly, this implies that the partition between these

variables is not referential-independent but agent-specific. Each agent has its way of partitioning (seeing) the environment. Also, the environment can be thought of as a mediator of many types of interactions, abstracting away potentially many other (behavior) agents who are also interacting with and have their own perspectives compared to the CRL agent, which is the one we care about. Fig. 4 illustrates this situation where the l.h.s. includes other agents interacting in the environment following their own behavior policies or natural variations, such as the laws of nature, including the wind, gravity, and natural selection. On the r.h.s., the CRL agent is depicted with its own views and capabilities.

2.2.1 OBSERVATIONAL DISTRIBUTIONS

The CRL agent has two primary ways of interacting with the system: by perceiving the world through its sensors (by "seeing") or by performing intervention through its actuators (by "doing"). This can be formalized through layers 1 and 2 of the PCH, which are shown in the middle of Fig. 4. When the CRL agent passively observes how events unfold in time (marked in blue), the variations and dynamics in the system come from the other behavior agents (in the left), which can be natural or artificial. Since the CRL agent does not know the other agents' perspectives (their views, goals, and policies), these other interactions (marked in red) are of unknown nature and considered passive observations. In this case, the CRL agent does not deliberately interfere with the underlying SCM at any point in time. Actions $\boldsymbol{X}$ attain their values under the control of the natural/behavior's regime.

For any SCM $\mathcal{M}$, the collection of structural functions $\mathscr{F}$ defines a mapping from the system's exogenous (unobserved) variables $\boldsymbol{U}$ to the endogenous (observed) variables $\boldsymbol{V}$. The distribution over the exogenous variables $P(\boldsymbol{U})$ induces a joint distribution over endogenous variables $\boldsymbol{V}$, $P(\boldsymbol{V})$, which is called the *observational distribution*.

Definition 2 (Observational Distribution (Bareinboim et al., 2020)) *A SCM $\mathcal{M} = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P(\boldsymbol{U}) \rangle$ defines a joint probability distribution $P(\boldsymbol{Y})$ such that for any $\boldsymbol{Y} \subseteq \boldsymbol{V}$,*

$$P(\boldsymbol{y}) = \sum_{\boldsymbol{u}} \mathbb{1}\{\boldsymbol{Y}(\boldsymbol{u}) = \boldsymbol{y}\} P(\boldsymbol{u}), \tag{6}$$

where $\boldsymbol{Y}(\boldsymbol{u})$ is the solution of $\boldsymbol{Y}$ after evaluating functions in $\mathscr{F}$ given $\boldsymbol{U} = \boldsymbol{u}$. ■

We will consistently use $P(\boldsymbol{Y}; \mathcal{M})$ to represent an observational distribution $P(\boldsymbol{Y})$ evaluated with restriction in an SCM $\mathcal{M}$. The input $\mathcal{M}$ is omitted when the SCM is obvious from the context. The observational distribution in Eq. 6 could be evaluated following the procedure:

1. For each situation $\boldsymbol{U} = \boldsymbol{u}$,[5] the environment evaluates the mechanisms in $\mathscr{F}$ following a topological order (i.e., any variable in the l.h.s of function f_i is evaluated after the ones in the r.h.s.), and

2. The probability mass $P(\boldsymbol{u})$ is accumulated for each realization $\boldsymbol{U} = \boldsymbol{u}$ consistent with the event $\boldsymbol{Y} = \boldsymbol{y}$.

We provide examples of this evaluation process below with the canonical RL models discussed earlier, where the CRL agent passively observes the unfolding of other agents interacting with $\mathcal{M}$.

5. Recall that each instantiation $\boldsymbol{U} = \boldsymbol{u}$ represents unobserved factors that generate variations within the environment of interest. In our context, this may represent an individual, a situation, or a state.

Example 3 (MAB, Observational Distribution) *Consider the MAB model $\mathcal{M}^*_{\text{MAB}}$ defined earlier in Eq. 3. By passively observing the behaviors of the physicians and collecting data, the CRL agent can estimate that the average recovery rate of each patient as*

$$\begin{aligned} P(Y=1) &= P(X=0,Y=1) + P(X=1,Y=1) && (7)\\ &= P(U \geq 0.8, U < 0.4) + P(U < 0.8, U < 0.4-\Delta) && (8)\\ &= P(U < 0.4-\Delta) && (9)\\ &= 0.4-\Delta. && (10) \end{aligned}$$

The agent has access only to the value in the r.h.s. of Eq. 10, and Nature does the evaluation process itself. In particular, patients are evaluated following the two 2-step procedure described in Def. 2, following the population's proportions and the corresponding structural mechanisms.

To further consider the effectiveness of the treatments, the CRL agent may collect data so that it can also compute the recovery rate for the specific treatment $X=0$. That is,

$$\begin{aligned} P(Y=1 \mid X=0) &= P(U < 0.4 \mid X=0) && (11)\\ &= P(U < 0.4 \mid U \geq 0.8) && (12)\\ &= 0 && (13) \end{aligned}$$

Similarly, the recovery rate conditioning on event $X=1$ is given by

$$\begin{aligned} P(Y=1 \mid X=1) &= P(U > 0.4-\alpha \mid X=1) && (14)\\ &= P(U > 0.4-\Delta \mid U < 0.8) && (15)\\ &= 0.5 - 1.25\Delta. && (16) \end{aligned}$$

Again, the CRL agent has access only to the r.h.s. of the equation through data coming from sampling. Since the coefficient $\Delta \in (0, 0.4)$, then $0.5 - 1.25\Delta > 0$, which implies that

$$P(Y=1 \mid X=0) < P(Y=1 \mid X=1). \tag{17}$$

This seems to suggest that treatment $X=1$ achieves a better recovery rate than treatment $X=0$.

The conclusion that follows from Eq. 17 seems to be at odds with the earlier analysis based on the full knowledge of the SCM and the specific mechanisms of how Y comes about (as shown in Eq. 3), which concluded that $X=0$ is the optimal treatment.

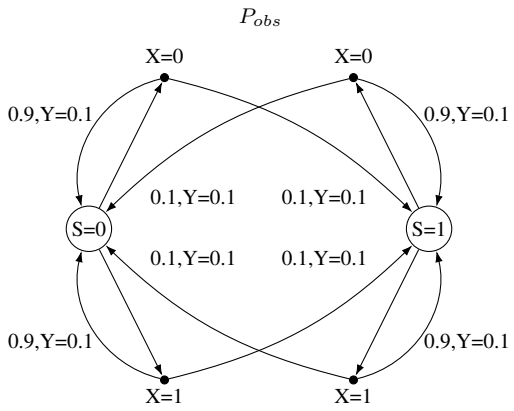
Note that this evaluation is based on passively collected data (from layer the PCH's $\mathscr{L}_1$), which means that a proper causal interpretation is generally not well-advised. The behavior agent may not be efficient in allocation or have different goals in mind. ■

Example 4 (MDP, Observational Distribution) *Consider the MDP $\mathcal{M}^*_{\text{MDP}}$ described in Eq. 5. We are interested in evaluating the inventory manager's cumulative profit $\mathbb{E}\left[\sum_{i=1}^{\infty} \gamma^{i-1} Y_i\right]$, where γ is a discount factor in the real interval $(0,1)$. Evaluating profit Y_i on day $i = 1, 2, \ldots$ in $\mathcal{M}^*_{\text{MDP}}$ gives:*

$$\begin{aligned} Y_i &= S_i \oplus X_i \oplus U_{i,1} \oplus U_{i,2} \\ &= S_i \oplus S_i \oplus U_{i,1} \oplus U_{i,1} \oplus U_{i,2} \\ &= U_{i,3} && (18) \end{aligned}$$

S_{i+1}	S_i	X_i	$P(s_{i+1} \mid s_i, x_i)$	$\mathbb{E}[Y_i \mid s_i, x_i]$
0	0	0	0.9	0.1
0	0	1	0.9	0.1
0	1	0	0.1	0.1
0	1	1	0.1	0.1
1	0	0	0.1	0.1
1	0	1	0.1	0.1
1	1	0	0.9	0.1
1	1	1	0.9	0.1

(a)



(b)

Figure 5: Observational distributions of the MDP model $\mathcal{M}^*_{\text{MDP}}$ described in Eq. 5

That is, the manager's inventory control policy generates an expected profit $\mathbb{E}[Y_i] = P(U_{i,2} = 1) = 0.1$ *every day. For concreteness, we consider* $\gamma = 0.9$*. The store's expected cumulative reward is:*

$$\mathbb{E}\left[\sum_{i=1}^{\infty} \gamma^{i-1} Y_i\right] = \sum_{i=1}^{\infty} \gamma^{i-1} \mathbb{E}[Y_i] = \frac{0.1}{1-\gamma} = 1. \tag{19}$$

We now consider the transition distribution $P(S_{i+1} \mid S_i, X_i)$ *and expected reward* $\mathbb{E}[Y_i \mid S_i, X_i]$ *conditional on decision* X_i *and inventory size* S_i *on day* i*. First, evaluating the inventory size* S_{i+1} *on the next day in* $\mathcal{M}^*_{\text{MDP}}$ *gives:*

$$P(S_{i+1} \mid S_i = s_i, X_i = x_i) = P((s_i \vee x_i) \oplus U_{i,1} \oplus U_{i,2} \mid S_i = s_i, X_i = x_i) \tag{20}$$

For $\mathcal{M}^*_{\text{MDP}}$ *defined in Eq. 5, observing events* $S_i = 0, X_i = 0$ *implies that* $U_{i,1} = 0$*, so*

$$\begin{aligned} P(S_{i+1} = 1 \mid S_i = 0, X_i = 0) &= P((0 \vee 0) \oplus 0 \oplus U_{i,2} = 1 \mid S_i = 0, X_i = 0) && (21)\\ &= P(U_{i,2} = 1) && (22)\\ &= 0.1. && (23) \end{aligned}$$

The second step holds since the stochastic demand $U_{i,2}$ *is an independent factor only affecting* S_{i+1}*. Similarly, evaluating the expected reward* Y_i *in* $\mathcal{M}^*_{\text{MDP}}$ *gives:*

$$\mathbb{E}[Y_i \mid S_i = s_i, X_i = x_i] = \mathbb{E}[s_i \oplus x_i \oplus U_{i,1} \oplus U_{i,3} \mid S_i = s_i, X_i = x_i] \tag{24}$$

For $S_i = 0, X_i = 0$*, the above equation could be further written as*

$$\begin{aligned} \mathbb{E}[Y_i \mid S_i = 0, X_i = 0] &= \mathbb{E}[0 \oplus 0 \oplus 0 \oplus U_{i,3}] && (25)\\ &= P(U_{i,3} = 1) && (26)\\ &= 0.1 && (27) \end{aligned}$$

The detailed parametrizations of the $\mathscr{L}_1$-distribution, $P(S_{i+1} \mid S_i, X_i)$ and $\mathbb{E}[Y_i \mid S_i, X_i]$, are described in Fig. 5a. Following the convention in reinforcement learning, this parametrization can be represented through a probabilistic finite-state machine as shown in Fig. 5b, which is read as follows. Assuming the current state is $S = 0$, there are two outgoing transitions to $X = 0$ and $X = 1$, which represent two possible actions the inventory manager could be observed to take. If the manager takes action $X = 0$, the transition of returning back to the state $S = 0$ has a probability of 0.9, and is associated with an average reward $Y = 0.1$, as indicated in the arrow. That is, the transition probability $P(S_{i+1} = 0 \mid S_i = 0, X_i = 0) = 0.9$, and the reward function $\mathbb{E}[Y_i \mid S_i = 0, X_i = 0] = 0.1$. On the other hand, it may also transition to state $S = 1$ with probability 0.1, resulting in a subsequent reward of $Y = 0.1$. That is, the transition function $P(S_{i+1} = 0 \mid S_i = 1, X_i = 0) = 0.1$, and the reward function $\mathbb{E}[Y_i \mid S_i = 1, X_i = 0] = 0.1$. ■

So far, we have defined a model for typical, template-like RL environments and some implications of passively observing such environments, including collecting data from the corresponding observational ($\mathscr{L}_1$) distributions.

2.2.2 INTERVENTIONAL DISTRIBUTIONS

In this section, we discuss how to model the agent's interventions in the world whenever they aim to bring some state of affairs about. In particular, we consider interventions that build on previous states' history and actions. These types of interventions are usually called *soft* or *policy interventions*, and we will follow the treatment developed in (Correa and Bareinboim, 2020a,b).[6]

Formally, we denote by $\pi_{\boldsymbol{X}}$ a sequence of decision rules $\{\pi_X \mid \forall X \in \boldsymbol{X}\}$ over an arbitrary set of action variables $\boldsymbol{X}$. Each π_X is a function that determines values of X taking some other variables $\boldsymbol{S}_X$ as input; that is, $X \leftarrow \pi_X(\boldsymbol{S}_X)$. With a slight abuse of notation, we denote it by $\pi_X(X \mid \boldsymbol{S}_X)$, the stochastic policy mapping from values of $\boldsymbol{S}_X$ to the probabilities space over domains of every $X \in \boldsymbol{X}$.[7] Such policies $\pi_{\boldsymbol{X}}$ are also referred to as adaptive treatment strategies or treatment policies in the healthcare literature (Murphy et al., 2001a; Chakraborty and Murphy, 2014). These decision rules provide an effective vehicle for personalized medicine for chronic conditions, in which treatment is repeatedly tailored to a patient's dynamic state.

A policy dictates the actions that an agent (e.g., a physician, an ad-placement engine) could take based on the values of the states that it observes in the underlying SCM. A policy intervention $\text{do}(\boldsymbol{X} \leftarrow \pi_{\boldsymbol{X}})$ following a policy $\pi_{\boldsymbol{X}}$ (for short, $\text{do}(\pi_{\boldsymbol{X}})$) is an operation that replaces the original behavior policy f_X associated with every variable $X \in \boldsymbol{X}$ with the corresponding decision rule π_X.[8] We formally define the new world that emerges when the current one is submitted to an intervention through the notion of a submodel:

Definition 3 (Submodel) *Let $\mathcal{M} = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P \rangle$ be an SCM and let $\pi_{\boldsymbol{X}}$ be a policy over actions $\boldsymbol{X} \subseteq \boldsymbol{V}$. A submodel $\mathcal{M}_{\pi_{\boldsymbol{X}}}$ of $\mathcal{M}$ is an SCM $\langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}_{\pi_{\boldsymbol{X}}}, P(\boldsymbol{U}) \rangle$ where*

$$\mathscr{F}_{\pi_{\boldsymbol{X}}} = \{f_V : \forall V \in \boldsymbol{V} \setminus \boldsymbol{X}\} \cup \{\pi_X : \forall X \in \boldsymbol{X}\}. \tag{28}$$

6. There is a growing literature interested in different features of these interventions, refer to (Pearl, 2000, Ch. 4) for some historical discussion, and also (Dawid, 2002; Didelez et al., 2006; Tian, 2008).
7. Formally, this means that the original pair, natural/behavior function f_X and exogenous term U_X is replaced with another pair, a new function f_X^* and an independent noise U_X^*, generating the purported conditional distribution π_X behavior. For simplicity, we ignore U_X^* and write a stochastic policy $X \sim \pi(x|\boldsymbol{s}_X)$.
8. This overwrites the natural/behavior policy and is, therefore, oblivious to how other agents, artificial or natural, were operating in the environment before the intervention.

■

In words, whenever the causal system $\mathcal{M}$ is submitted to an intervention $\pi_{\boldsymbol{X}}$, the equations relative to the original mechanisms (whatever these were) are replaced with the ones corresponding to the intervention, and all other equations remain the same; the resultant model is called $M_{\pi_{\boldsymbol{X}}}$. A significant special class of interventions is called atomic, $\text{do}(\boldsymbol{X} \leftarrow \boldsymbol{x})$ (for short, $\text{do}(\boldsymbol{x})$), which sets the values of variables $\boldsymbol{X}$ to some constants $\boldsymbol{x}$. That is, decision rules are defined as $X \leftarrow x$ for every variable $X \in \boldsymbol{X}$.[9] The submodel induced by an atomic intervention $\text{do}(\boldsymbol{x})$ in an SCM $\mathcal{M}$ is usually written as $\mathcal{M}_{\boldsymbol{x}}$ (Pearl, 2000, p. 204). Further note that this is a derived model defined over the original SCM $\mathcal{M}$. In other words, the SCM is generative of multiple worlds, while $\mathcal{M}_{\boldsymbol{x}}$ represents one of these worlds.[10]

With the context of the sequential decision-making setting in mind, we provide a few concrete examples below demonstrating the evaluation of policies in canonical RL environments.

Example 5 (MAB, Submodel) *Let us consider again the MAB model $\mathcal{M}^*_{\text{MAB}}$ in Eq. 3. Due to new HIPAA privacy rules, the age of patients (U) is protected and cannot be disclosed, meaning they are unobserved from the CRL agent's perspective. Consider a policy $\pi \triangleq X \leftarrow x$ that sets treatment X to a constant $x \in \{0, 1\}$. The submodel $\mathcal{M}^*_{\text{MAB}_x}$ induced by atomic intervention do$(X \leftarrow x)$ is a tuple*

$$\mathcal{M}^*_{\text{MAB}_x} = \langle \boldsymbol{U} = \{U\}, \boldsymbol{V} = \{X, Y\}, \mathscr{F}_x, P(U) \rangle, \tag{29}$$

where

$$\mathscr{F}_x = \begin{cases} X \leftarrow x, \\ Y \leftarrow \mathbb{1}\{U < 0.4 - \Delta X\} \end{cases} \tag{30}$$

More broadly, a stochastic policy $\pi(X)$ is a probability distribution over domains of arm choice $X \in \{0, 1\}$. The submodel entailed by intervention do$(X \sim \pi(X))$ is a tuple

$$\mathcal{M}^*_{\text{MAB}_\pi} = \langle \boldsymbol{U} = \{U\}, \boldsymbol{V} = \{X, Y\}, \mathscr{F}_\pi, P(U) \rangle, \tag{31}$$

where

$$\mathscr{F}_\pi = \begin{cases} X \sim \texttt{Bernoulli}(\pi(X = 1)), \\ Y \leftarrow \mathbb{1}\{U < 0.4 - \Delta X\} \end{cases} \tag{32}$$

In words, the CRL agent following a stochastic policy $\pi(X)$ prescribes treatment $X = x$ with probability $\pi(X = x)$ where $x \in \{0, 1\}$. The patient's age (U) is not disclosed to the CRL agent and thus is not considered when choosing the treatment at the decision-making time. ■

9. This basic primitive has appeared at different times and contexts through causality's history. It was introduced in econometrics by Haavelmo (1943); Strotz and Wold (1960). In statistics, potential outcomes were introduced in the context of randomized experiments by Neyman (1923) and then connected with observational studies by Rubin (1974). In mathematical logic, counterfactuals were discussed by Lewis (1973) with possible worlds semantics. Pearl developed a general and algorithmic treatment through graphical models in AI (Pearl, 1993, 1995) .

10. The importance of this notion comes from the fact that it will allow us to represent the idea of causal effect, which is critical in evaluating the effect of actions. Of course, this is a semantical definition and operationalizing it in practice, whenever $\mathcal{M}$ is unknown, will be part of the inferential challenge faced by the CRL agent.

Example 6 (Markov Decision Process's Submodel) *Consider the MDP $\mathcal{M}^*_{\text{MDP}}$ in Eq. 5 and an atomic intervention do$(X_1 \leftarrow x_1, \dots, X_i \leftarrow x_i)$, where the stocking decision X_i for every day i is fixed at constant $x_i = 0, 1$. The induced submodel is described by the tuple*

$$\mathcal{M}_{\text{MDP}_{\boldsymbol{x}}} = \langle \boldsymbol{U} = \{U_{i,1}, U_{i,2}, U_{i,3}\}, \boldsymbol{V} = \{X_i, Y_i, S_i\}, \mathscr{F}_{\boldsymbol{x}} = \{\mathscr{F}_{x_i}\}, P(\boldsymbol{U})\rangle_{i=1,2,\dots} \tag{33}$$

and $\mathscr{F}_{x_i}$ is the post-interventional system dynamics from day $i-1$ to day i given by

$$\mathscr{F}_{x_i} = \begin{cases} S_i \leftarrow (S_{i-1} \vee X_{i-1}) \oplus U_{i-1,1} \oplus U_{i-1,2}, \\ X_i \leftarrow x_i \\ Y_i \leftarrow S_i \oplus X_i \oplus U_{i,1} \oplus U_{i,3} \end{cases} \tag{34}$$

To improve the long-term profit, the store decides to automate the day-to-day inventory control using the CRL agent. Since operation errors $U_{i,1}$ and fluctuations in demands $U_{i,2}$ and pricing $U_{i,3}$ are not recorded in the store's system, they are unobserved from the CRL agent's perspective.

A policy π in the MDP model $\mathcal{M}$ is a sequence of decision rules $\pi = (\pi_1, \pi_2, \dots)$, one for every time step i. Every decision rule $\pi_i(X_i \mid S_i)$ is a conditional distribution mapping from the domain of state S_i to action X_i, for every step $i = 1, 2, \dots$. The submodel entailed by intervention do$(X_1 \sim \pi_1(X_1 \mid S_1), X_2 \sim \pi_2(X_2 \mid S_2), \dots)$ is described by the tuple

$$\mathcal{M}^*_{\text{MDP}_\pi} = \langle \boldsymbol{U} = \{U_{i,1}, U_{i,2}, U_{i,3}\}, \boldsymbol{V} = \{X_i, Y_i, S_i\}, \mathscr{F}_\pi = \{\mathscr{F}_{\pi_i}\}, P(\boldsymbol{U})\rangle_{i=1,2,\dots}, \tag{35}$$

where the causal mechanisms $\mathscr{F}_{\pi_i}$ transitioning from day $i-1$ to i is given by

$$\mathscr{F}_{\pi_i} = \begin{cases} S_i \leftarrow (S_{i-1} \vee X_{i-1}) \oplus U_{i-1,1} \oplus U_{i-1,2}, \\ X_i \sim \texttt{Bernoulli}(\pi_i(X_i = 1 \mid S_i)) \\ Y_i \leftarrow S_i \oplus X_i \oplus U_{i,1} \oplus U_{i,3} \end{cases} \tag{36}$$

In words, the CRL agent following a stochastic policy $\pi = (\pi_1, \pi_2, \dots)$ decides whether to restock $X_i = 1$ with probability $\pi_i(X_i = 1 \mid S_i)$ for every time step $i = 1, 2, \dots$. The agent's decision is free from operation errors and thus is not affected by exogenous variables $U_{i,1}, U_{i,2}, U_{i,3}$. ■

The usefulness of a submodel is that it gives semantics to how reality will behave when submitted to a new interventional condition. Similar to the observational distribution (Def. 2), an SCM also gives a natural valuation of consequences induced by interventions do$(\pi_{\boldsymbol{X}})$ following a policy $\pi_{\boldsymbol{X}}$. The impact of the intervention on reward signals $\boldsymbol{Y}$ is called a potential response. The definition of potential response of atomic intervention do$(\boldsymbol{x})$ was provided in (Pearl, 2000, Def. 7.1.4), and next, we introduce a generalization to policy interventions do$(\pi_{\boldsymbol{X}})$.

Definition 4 (Potential Response) *An SCM $\mathcal{M} = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P\rangle$, let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be two sets of variables in $\boldsymbol{V}$, $\pi_{\boldsymbol{X}}$ be a policy over $\boldsymbol{X}$, and $\boldsymbol{u}$ be a unit. The potential response $\boldsymbol{Y}_{\pi_{\boldsymbol{X}}}(\boldsymbol{u})$ is defined as the solution for $\boldsymbol{Y}$ of the set of equations $\mathscr{F}_{\pi_{\boldsymbol{X}}}$ in $\mathcal{M}$. That is, $\boldsymbol{Y}_{\pi_{\boldsymbol{X}}}(\boldsymbol{u}) \triangleq \boldsymbol{Y}(\boldsymbol{u}; \mathcal{M}_{\pi_{\boldsymbol{X}}})$.*

Formally, the interventional distributions produced by do$(\pi_{\boldsymbol{X}})$ in an SCM $\mathcal{M}$ are distributions over endogenous variables in submodel $\mathcal{M}_{\pi_{\boldsymbol{X}}}$.

Definition 5 (Interventional Distribution) *An SCM $\mathcal{M} = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P(\boldsymbol{U}) \rangle$ induces a family of joint probability distributions over $\boldsymbol{V}$, one for each intervention do$(\pi_{\boldsymbol{X}})$, where $\pi_{\boldsymbol{X}}$ is a policy over actions $\boldsymbol{X} \subseteq \boldsymbol{V}$. For each endogenous set $\boldsymbol{Y} \subseteq \boldsymbol{V}$,*

$$P_{\pi_{\boldsymbol{X}}}(\boldsymbol{Y}) \equiv \sum_{\boldsymbol{u}} \{\boldsymbol{Y}_{\pi_{\boldsymbol{X}}}(\boldsymbol{u}) = \boldsymbol{y}\} P(\boldsymbol{u}), \tag{37}$$

where $\boldsymbol{Y}_{\pi_{\boldsymbol{X}}}(\boldsymbol{u})$ is the potential response of intervention do$(\pi_{\boldsymbol{X}})$ on variables $\boldsymbol{Y}$ (Def. 4).

Distributions entailed by policy interventions do$(\pi_{\boldsymbol{X}})$, defined in Eq. 37, could be evaluated using a procedure described as follows.

1. Replace the mechanism of each $X \in \boldsymbol{X}$ with the corresponding functions π_X generating functions $\mathscr{F}_{\pi_{\boldsymbol{X}}}$ (Eq. 28), which induces a submodel $\mathcal{M}_{\pi_{\boldsymbol{X}}}$ (of $\mathcal{M}$);
2. For each situation $\boldsymbol{U} = \boldsymbol{u}$, the environment evaluates $\mathscr{F}_{\pi_{\boldsymbol{X}}}$ following a valid order (where any variable in the l.h.s. is evaluated after the ones in the r.h.s.), and
3. The probability mass $P(\boldsymbol{U} = \boldsymbol{u})$ is then accumulated for each instantiation $\boldsymbol{U} = \boldsymbol{u}$ consistent with the event $\boldsymbol{Y} = \boldsymbol{y}$ in the submodel $\mathcal{M}_{\pi_{\boldsymbol{X}}}$.

As a special case, we denote by $P_{\boldsymbol{x}}(\boldsymbol{Y})$ the interventional distribution entailed by atomic interventions do$(\boldsymbol{x})$, which is a joint distribution over variables Y in submodel $\mathcal{M}_{\boldsymbol{x}}$. The following examples demonstrate the evaluation of both atomic and policy interventional distributions in some canonical RL environments.

Example 7 (MAB, Interventional Distribution, Atomic) *For the MAB model $\mathcal{M}^*_{\text{MAB}}$ described in Eq. 3, we compute the interventional distribution $P(Y \mid do(X \leftarrow x))$, $x \in \{0, 1\}$. Evaluating the recovery of the patient Y in submodel $M^*_{\text{MAB}_{X \leftarrow 0}}$ described in Eq. 30 gives*

$$\begin{aligned} P_{X \leftarrow 0}(Y = 1) &= P(U < 0.4) && (38)\\ &= 0.4 && (39) \end{aligned}$$

Similarly, the patient's recovery rate of treatment $X \leftarrow 0$ is equal to

$$\begin{aligned} P_{X \leftarrow 1}(Y = 1) &= P(U < 0.4 - \Delta) && (40)\\ &= 0.4 - \Delta. && (41) \end{aligned}$$

Since the coefficient $\Delta > 0$,

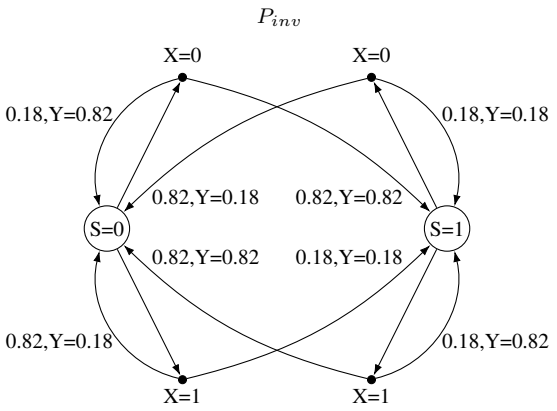
$$P_{X \leftarrow 0}(Y = 1) > P_{X \leftarrow 1}(Y = 1). \tag{42}$$

A few observations follow. First, a policy prescribing treatment $X \leftarrow 0$ implies a higher chance of patients' recovery compared to $X \leftarrow 1$, i.e., $X \leftarrow 0$ is the optimal treatment. Second, this matches the analysis based on the true mechanism of Y as described in Eq. 3. ■

A significant challenge arises since the agent doesn't have access to the true description of the environment, encoded as the SCM $\mathcal{M}$, it will need to perform interventions physically and set action $X = x$, despite the other factors, to obtain samples from the two distributions described by Eqs. 39 and 41. In turn, we discuss more complex interventions from a non-atomic policy.

S_{i+1}	S_i	X_i	$P_{x_i}(s_{i+1} \mid s_i)$	$\mathbb{E}_{x_i}[Y_i \mid s_i]$
0	0	0	0.18	0.82
0	0	1	0.82	0.18
0	1	0	0.82	0.18
0	1	1	0.82	0.82
1	0	0	0.82	0.82
1	0	1	0.18	0.18
1	1	0	0.18	0.18
1	1	1	0.18	0.82

(a)



(b)

Figure 6: Interventional distributions of the MDP model described in Eq. 5

Example 8 (MAB, Interventional Distribution, Policy) *For the same MAB environment $\mathcal{M}^*_{\text{MAB}}$ described in the previous example (and Eq. 3), we have that event $X = 1$ is equivalent to $U < 0.5$. The observational distribution can be obtained by using Eq. 6 and $\mathcal{M}^*_{\text{MAB}}$, which leads to $P(X = 1) = 0.5$. In words, the physician seems to be prescribing treatment X uniformly at random. Recall that the physician's recovery rate based on Eq. 10 was $P(Y = 1) = 0.5 + \alpha$.*

One may be tempted to surmise that the CRL agent could achieve the same performance as the physician by 'cloning' its random policy, i.e.,

$$\pi(X = x) = P(X = x) = 0.5. \tag{43}$$

Perhaps surprisingly, this is not the case. Specifically, evaluating the recovery Y in submodel $\mathcal{M}^*_{\text{MAB}_\pi}$ described in Eq. 30, gives:

$$P_\pi(Y = 1) = \sum_{x=0,1} \pi(x) P_{X \leftarrow x}(Y = 1) \tag{44}$$

$$= 0.5 + 0.5\alpha. \tag{45}$$

Whenever coefficient $\alpha > 0$, the recovery rate obtained by the CRL agent given by Eq. 45 will be smaller than the physician's (Eq. 10). ■

This example highlights the clear difference between observational (seeing) and interventional (doing) distributions from the perspective of the CRL agent. Also, when the behavior and the CRL agents have different perceptual capabilities and views of the environment, naively copying (or cloning) the nominal behavior policy $P(X)$ may not lead to a successful policy (as shown by Ex. 7). There is no formal basis to pursue such a strategy since these two distributions are different, and the empirical gap could be significant.[11]

11. This leads to a more refined discussion of imitation learning, which is provided in Sec. 8 and accompanied by proper causal modeling and solutions.

Example 9 (MDP, Interventional Distribution, Atomic) *Consider the MDP environment $\mathcal{M}^*_{\text{MDP}}$ described in Eq. 5 and the transition distribution $P_{x_i}(S_{i+1} \mid S_i)$ and expected reward $\mathbb{E}_{x_i}[Y_i \mid S_i]$ induced by performing intervention $do(X_i \leftarrow x_i)$ at inventory size S_i on day i. Evaluating the inventory size S_{i+1} on the next day in MDP $\mathcal{M}^*_{\text{MDP}_{\boldsymbol{x}}}$ (Eq. 34) gives*

$$P_{X_i \leftarrow x_i}(S_{i+1} \mid S_i = s_i) = P((s_i \vee x_i) \oplus U_{i,1} \oplus U_{i,2} \mid S_i = s_i) \tag{46}$$
$$= P((s_i \vee x_i) \oplus U_{i,1} \oplus U_{i,2}) \tag{47}$$

The second step holds since the stochastic demand $U_{i,2}$ is an independent variable only affecting S_{i+1}. For $S_i = 0, X_i \leftarrow 0$, the above equation could be written as

$$P_{X_i \leftarrow 0}(S_{i+1} = 1 \mid S_i = 0) = P((0 \vee 0) \oplus U_{i,1} \oplus U_{i,2} = 1) \tag{48}$$
$$= P(U_{i,1} \oplus U_{i,2} = 1) \tag{49}$$
$$= 0.82. \tag{50}$$

Similarly, evaluating the expected reward Y_i in MDP $\mathcal{M}^*_{\text{MDP}_{\boldsymbol{x}}}$ gives:

$$\mathbb{E}_{X_i \leftarrow x_i}[Y_i \mid S_i = s_i] = \mathbb{E}[s_i \oplus x_i \oplus U_{i,1} \oplus U_{i,3} \mid S_i = s_i] \tag{51}$$
$$= \mathbb{E}[s_i \oplus x_i \oplus U_{i,1} \oplus U_{i,3}] \tag{52}$$

For $S_i = 0, X_i \leftarrow 0$, the company's expected profit is equal to

$$\mathbb{E}_{X_i \leftarrow 0}[Y_i \mid S_i = 0] = \mathbb{E}[0 \oplus 0 \oplus U_{i,1} \oplus U_{i,3}] \tag{53}$$
$$= P(U_{i,1} \oplus U_{i,3} = 1) \tag{54}$$
$$= 0.82 \tag{55}$$

The complete parametrizations of $P_{X_i \leftarrow x_i}(S_{i+1} \mid S_i = s_i)$ and $\mathbb{E}_{X_i \leftarrow x_i}[Y_i \mid S_i = s_i]$ are shown in Fig. 6a. The dynamic process described in Fig. 6b shows a compact graphical representation of this parametrization where transition probabilities $\mathcal{T}$ and reward function $\mathcal{R}$ are given by

$$\mathcal{T}(s, x, s') = P_{X_i \leftarrow x}\left(S_{i+1} = s' \mid S_i = s\right) \tag{56}$$
$$\mathcal{R}(s, x) = \mathbb{E}_{X_i \leftarrow x}[Y_i \mid S_i = s] \tag{57}$$

Note that compared with probabilities in Table 5a, interventional distributions $P_{X_i}(S_{i+1} \mid S_i)$, $\mathbb{E}_{X_i}[Y_i \mid S_i]$ and observational distributions $P(S_{i+1} \mid S_i, X_i)$, $\mathbb{E}[Y_i \mid S_i, X_i)]$ do not coincide in the MDP model $\mathcal{M}^*_{\text{MDP}}$ described in Eq. 5. ■

Following Examples 4 and 9, we note that the standard definition of MDP found in the literature (Puterman, 1994; Sutton and Barto, 1998) based on a specific pair of transition probabilities $\mathcal{T}$ and reward function $\mathcal{R}$ only provides an abstraction for distributions in a single layer of the PCH, either observational or interventional. On the other hand, the SCM description of the MDP environment (e.g., Eq. 5) provides a complete specification that allows for inferences across all layers of the PCH. This observation is interesting considering the distinct nature of the PCH's layers and each of the type of distributions associated with each of them. We will elaborate on this further in Sec. 3.3.

We can also infer about the effects of non-atomic interventions following a Markov policy π which determines values of every action X_i based on the observed state S_i at every decision horizon $i = 1, 2, \ldots$. Our next example demonstrates such complex interventions in MDP environments.

Example 10 (MDP, Interventional Distribution, Policy) *Consider the MDP $\mathcal{M}^*_{\text{MDP}}$ described in Eq. 5 and a policy $\pi = (\pi_1(X_1 \mid S_1), \pi_2(X_2 \mid S_2), \dots)$. Evaluating the transition distribution from the current state S_i to next state S_{i+1} in submodel $\mathcal{M}^*_{\text{MDP}_\pi}$ (Eq. 36) gives*

$$P_\pi\left(S_{i+1} \mid S_i = s_i, X_i = x_i\right) = P\left((s_i \vee x_i) \oplus U_{i,1} \oplus U_{i,2} \mid S_i = s_i, X_i = x_i\right) \tag{58}$$
$$= P\left((s_i \vee x_i) \oplus U_{i,1} \oplus U_{i,2}\right) \tag{59}$$

The second step holds since the exogenous variable $U_{i,2}$ is an independent variable only affecting the next state S_{i+1}. It follows from the evaluation in Eq. 47 that the transition distribution remains invariant across atomic and policy interventions. That is,

$$P_\pi\left(S_{i+1} \mid S_i = s_i, X_i = x_i\right) = P_{X_i = x_i}\left(S_{i+1} \mid S_i = s_i\right) \tag{60}$$

Similarly, evaluating the expected reward Y_i conditioning on current state S_i and observed action X_i in submodel $\mathcal{M}^*_{\text{MDP}_\pi}$ gives:

$$\mathbb{E}_\pi\left[Y_i \mid S_i = s_i, X_i = x_i\right] = \mathbb{E}\left[s_i \oplus x_i \oplus U_{i,1} \oplus U_{i,3} \mid S_i = s_i, X_i = x_i\right] \tag{61}$$
$$= \mathbb{E}\left[s_i \oplus x_i \oplus U_{i,1} \oplus U_{i,3}\right] \tag{62}$$

Again, it follows from Eq. 52 that the conditional reward function remains the same for both atomic and policy interventions:

$$\mathbb{E}_\pi\left[Y_i \mid S_i = s_i, X_i = x_i\right] = \mathbb{E}_{X_i \leftarrow x_i}\left[Y_i \mid S_i = s_i\right] \tag{63}$$

More specifically, let policy π be defined such that $X_i \leftarrow S_i$ for every time step $i = 1, 2, \dots$. Evaluating the profit Y_i on day i evokes submodel $\mathcal{M}^*_{\text{MDP}_\pi}$ and is given by:

$$\mathbb{E}_\pi\left[Y_i\right] = \mathbb{E}_{X_i \leftarrow S_i}\left[S_i \oplus X_i \oplus U_{i,1} \oplus U_{i,3}\right] \tag{64}$$
$$= \mathbb{E}\left[S_i \oplus S_i \oplus U_{i,1} \oplus U_{i,3}\right] \tag{65}$$
$$= 0.82 \tag{66}$$

The cumulative profit induced by policy π with discount factor $\gamma = 0.9$ is then equal to:

$$\mathbb{E}_\pi\left[\sum_{i=1}^{\infty} \gamma^{i-1} Y_i\right] = \sum_{i=1}^{\infty} \gamma^{i-1} \mathbb{E}_\pi\left[Y_i\right] \tag{67}$$
$$= \frac{0.82}{1-\gamma} \tag{68}$$

Computing the above equation gives $\mathbb{E}_\pi\left[\sum_{i=1}^{\infty} \gamma^{i-1} Y_i\right] = 8.2$, which outperforms the inventory manager's performance (behavior agent) $\mathbb{E}\left[\sum_{i=1}^{\infty} \gamma^{i-1} Y_i\right] = 1$ (Eq. 19) under the behavior policy. In other words, it is more profitable to replace the manager with the CRL agent and automate the inventory management process in this case. ■

2.3 Encoding Structural Assumptions through Causal Diagrams

Even though SCMs are well-defined and provide precise semantics to the various types of distributions underlying the PCH, as discussed previously, one critical observation is that, in practice, they

are usually not observable by the CRL agent. Further assumptions are needed if one wants to reason about the underlying SCM. We will introduce an object called a *causal diagram* to encode assumptions about this SCM. There are different ways a causal diagram can be specified, including (1) as a template model (e.g., MABs, MDPs), (2) through prior knowledge, or (3) through a structural learning algorithm. We first describe the semantics of this object and a construction procedure that allows one to systematically articulate this causal diagram from a coarse, qualitative understanding of the underlying SCM.

Definition 6 (Causal Diagram (Pearl, 2000; Bareinboim et al., 2020)) *Consider the SCM* $\mathcal{M} = \langle \boldsymbol{V}, \boldsymbol{U}, \mathscr{F}, P(\boldsymbol{U}) \rangle$. *A graph* $\mathcal{G}$ *is said to be a* causal diagram *(of* $\mathcal{M}$*) if:*

1. *there is a vertex for every endogenous variable* $V_i \in \boldsymbol{V}$,
2. *there is an edge* $V_i \to V_j$ *if* V_i *appears as an argument of the mechanism* $f_j \in \mathscr{F}$,
3. *there is a bidirected edge* $V_i \dashleftarrow\!\dashrightarrow V_j$ *if the corresponding* $U_i, U_j \subset \boldsymbol{U}$ *are correlated or the corresponding functions* f_i, f_j *share some* $U_{ij} \in \boldsymbol{U}$ *as an argument.* ■

In words, there is an edge from endogenous variables V_i to V_j whenever V_j "listens to"[12] V_i for determining its value. Similarly, a bidirected edge between V_i and V_j indicates shared, unobserved information affecting how both, or whether V_i and V_j "listens" to the same source of exogenous variations. Note that while the SCM contains explicit, quantitative information about all structural mechanisms ($\mathscr{F}$) and exogenous probability distribution ($P(\boldsymbol{U})$), in contrast, the causal diagram encodes only *qualitative* information about which arguments were possibly used as inputs to the functions (from $\mathscr{F}$) and how the exogenous variations are related (from $P(\boldsymbol{U})$). The diagram abstracts out the specifics of the mechanisms $\mathscr{F}$ and distribution $P(\boldsymbol{U})$, retaining qualitative information about their possible arguments and independence structure, respectively.[13]

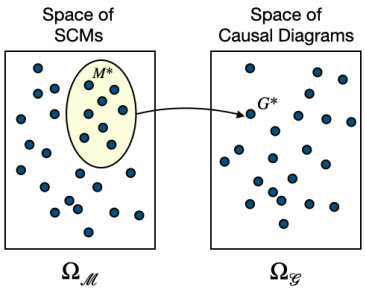


Figure 7: The space of SCMs/Causal diagrams are shown on the left/right side. The true SCM $\mathcal{M}^*$ and the corresponding causal diagram $\mathcal{G}^*$ are explicitly shown. The yellow area represents the subspace where these other SCMs generate the same $\mathcal{G}^*$.

I. Template models (knowledge-based). In practice, the CRL agent will not have access to the fully specified SCM and will operate based on the assumptions encoded in the given causal diagram. This will represent a major inferential challenge since the diagram is much weaker than the SCM, and there are various SCMs (marked in yellow in Fig. 7) equally compatible with the same causal

12. This construction lies at the heart of the type of knowledge causal models represent, as suggested in (Pearl and Mackenzie, 2018, pp. 129): "This listening metaphor encapsulates the entire knowledge that a causal network conveys; the rest can be derived, sometimes by leveraging data;" for technical details, (Bareinboim et al., 2020, Sec. 1.4).
13. Furthermore, the existence of a directed arrow, e.g., $V_i \to V_j$, encodes the *possibility* of the mechanism of V_j to listen to variable V_i, but not its necessity. In this sense, the edges are non-committal; for instance, f_j may decide not to consider V_i's value. More formally, the assumptions are not encoded in the arrows in the diagram but in the missing arrows; each missing arrow ascertains that one variable is *certainly* not the argument of the other or that one exogenous source of variation is not correlated to another. The same idea is true regarding the bidirected arrows and the possibility of covariation of some unobserved factors.

diagram. Still, whenever these inferences are allowed, this will translate into different gains in decision-making precision and efficiency.

To ground this particular construction, we will consider the first way of specifying knowledge of the underlying SCM through pre-specified template models, as illustrated in the following example.

Example 11 (MAB's Causal Diagram) *Consider a template family of MAB models $\mathbb{M}_{\text{MAB}}$ which consists of SCMs $\mathcal{M}_{\text{MAB}}$ described by a tuple*

$$\mathcal{M}_{\text{MAB}} = \langle \boldsymbol{U} = \{U\}, \boldsymbol{V} = \{X, Y\}, \mathscr{F}, P(U) \rangle, \tag{69}$$

The causal mechanisms $\mathscr{F}$ are structural functions of the form:

$$\mathscr{F} = \begin{cases} X \leftarrow f_X(U), \\ Y \leftarrow f_Y(X, U) \end{cases} \tag{70}$$

To apply the graphical construction dictated by Def 6, the AI engineer starts the modeling process by examining each of the endogenous variables $\boldsymbol{V} = \{X, Y\}$, and adding them as nodes in the causal diagram. The corresponding diagram is illustrated in Fig. 8a and will be called $\mathcal{G}_{\text{MAB}}$.

They then consider the second and third conditions of Def. 6. The mechanism underlying the context variable can be written as,

$$X \leftarrow f_X(U), \tag{71}$$

which suggests that the action is determined by an exogenous variable U (in the natural regime). This is regardless of the specific form, f_X, of how these variables are realized in reality. The engineer may, in turn, think about the reward function, namely,

$$Y \leftarrow f_Y(X, U). \tag{72}$$

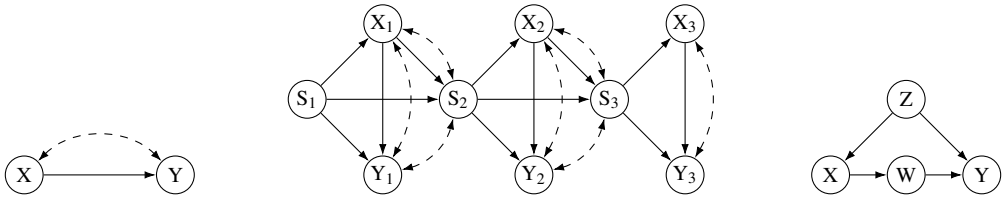
Eq. 72 suggests how, in reality, the rewards that come about may be influenced by the action X. Graphically, this is represented through the arrow $X \rightarrow Y$. Furthermore, since the mechanisms f_X and f_Y share the exogenous variable U, a bidirected arrow $X \dashleftarrow\!\dashrightarrow Y$ is added to $\mathcal{G}_{\text{MAB}}$.

Consider now a detailed MAB environment $\mathcal{M}^*_{\text{MAB}}$ defined in Eq. 3. Since $\mathcal{M}^*_{\text{MAB}}$ belongs to the template family $\mathbb{M}_{\text{MAB}}$, we could conclude that $\mathcal{G}_{\text{MAB}}$ is a causal diagram associated with $\mathcal{M}^*_{\text{MAB}}$. Note that this construction contrasts sharply with how detailed knowledge is encoded in the true SCM $\mathcal{M}^*_{\text{MAB}}$, as delineated in Def. 6. Interestingly enough, an entirely different functional form of the reward mechanism, say

$$Y \leftarrow \mathbb{1}\{U < 0.4 + \Delta X\} \tag{73}$$

would be equally compatible with the causal diagram depicted in Fig. 8a. Compared with the original reward in Eq. 3, the coefficient of X is flipped to Δ. This means it is preferable to pull arm $X \leftarrow 1$, which is the opposite of the optimal choice $X \leftarrow 0$ in the original model. ■

Similar construction and argument can be used when considering an MDP environment described in Eq. 5. The causal diagram in Fig. 8b is called $\mathcal{G}_{\text{MDP}}$ and represents causal relationships among variables shared across a template family of MDP environments. In the same way as the template causal diagram for MABs, this diagram $\mathcal{G}_{\text{MDP}}$ is non-committal regarding the form of the mechanisms $\mathscr{F}$ and the parametrization of the exogenous distribution $P(\boldsymbol{U})$.



(a) (b) (c)

Figure 8: Causal diagrams for (a) a multi-armed bandit (MAB); (b) a Markov decision process (MDP); and (c) an SCM representing a refinement of the MAB environment.

II. General causal models (knowledge-based). The graphical models discussed so far based on templates conveniently encapsulate structural information about the state, decision, and outcome variables.[14] As will become apparent in the following sections, this will naturally incur a cost in many practical CRL tasks. For now, we note that in some settings, additional information may be available that could be leveraged by the CRL agent.

For concreteness, consider the diagram in Fig. 8c. One way of thinking about it is as a refinement of the MAB diagram shown in Fig. 8a, where a confounder Z and a mediator W are now explicit. One natural question is how to test a model designed by the AI engineer. Interestingly enough, there are constraints imprinted by the SCM over the observational distribution $P(\boldsymbol{V})$, as well as the other PCH's distribution, that will allow the CRL agent to check whether its current working hypothesis is plausible, regardless of the idiosyncrasies of the properties of the distribution of exogenous $P(\boldsymbol{U})$ and the causal mechanisms $\mathscr{F}$ (e.g., monotonicity, linearity, separability).

We will discuss for now constraints known as *conditional independences* accompanied with a criterion known as *d*-separation (Pearl, 2000) that allows us to read such constraints from the model. A path p from a node X to a node Y in $\mathcal{G}$ is a sequence of edges that does not include a particular node more than once. It may go either along or against the direction of the edges. A path consisting of only bidirected edges is called a bidirected path. Formally, it goes as follows.

Definition 7 (*d*-separation (Pearl, 2000)) *A set $\boldsymbol{Z} \subseteq \boldsymbol{V}$ is said to block a path p in $\mathcal{G}$ if either*

1. *p contains at least one arrow-emitting node that is in $\boldsymbol{Z}$, or*
2. *p contains at least one collision node outside $\boldsymbol{Z}$ and has no descendant in $\boldsymbol{Z}$.*

If $\boldsymbol{Z}$ blocks all *paths from set $\boldsymbol{X}$ to set $\boldsymbol{Y}$, it is said to "d-separate $\boldsymbol{X}$ and $\boldsymbol{Y}$."*[15] ■

Before discussing some examples, we state one of the main results that connect d-separation statements made over the diagram $\mathcal{G}$ with constraints observed in the distribution $P(\boldsymbol{V})$.

Theorem 1 (Probabilistic Implications of *d*-Separation (Pearl, 2000)) *If $\boldsymbol{X}, \boldsymbol{Y}$ are* d*-separated by $\boldsymbol{Z}$ in a causal diagram $\mathcal{G}$, then $\boldsymbol{X}$ is independent of $\boldsymbol{Y}$ conditional on $\boldsymbol{Z}$ in every distribution P compatible with $\mathcal{G}$. Conversely, if $\boldsymbol{X}$ and $\boldsymbol{Y}$ are not* d*-separated by $\boldsymbol{Z}$ in a diagram $\mathcal{G}$, then $\boldsymbol{X}$ and $\boldsymbol{Y}$ are dependent conditional on $\boldsymbol{Z}$ in at least one distribution P compatible with $\mathcal{G}$.* ■

14. In reality, this class of causal diagrams was introduced under the rubric of *clustered diagrams* and different properties investigated, we refer readers to (Anand et al., 2021) for further details.
15. See Hayduk et al. (2003), Mulaik (2009), and Pearl (2009, pp. 335) for a gentle introduction to *d*-separation.

To illustrate these results, consider the causal diagram $\mathcal{G}$ in Fig. 8c and whether X and Y are d-separated. Note that there are two paths from X to Y,

$$p_1 : X \leftarrow \boxed{Z} \rightarrow Y, \tag{74}$$

$$p_2 : X \rightarrow \boxed{W} \rightarrow Y. \tag{75}$$

Since $\boldsymbol{Z} = \{\}$ in this case, both paths p_1 and p_2 are opened. One interpretation of this result is that there is a flow of information from X that is transmitted through Z, W that affects Y. Now, let's consider the separation statement when the conditioning set $\boldsymbol{Z} = \{Z, W\}$, or the confounder and mediator are in $\boldsymbol{Z}$. The first condition of the criterion is then immediately satisfied, and the paths p_1 and p_2 are "blocked." We can see through Thm. 1 that the following independence holds in $P(\boldsymbol{V})$:

$$(Y \perp\!\!\!\perp X \mid \{Z, W\}) . \tag{76}$$

Intuitively, once the values of Z and W are known, there is no information about X that will affect the likelihood of Y through p_1 and p_2, respectively. Readers are invited to check that the criterion is not satisfied if any intermediate variables are removed from the conditioning set.

III. General causal models (learning-based). Based on the marks imprinted by $\mathcal{M}^*$ on $P(\boldsymbol{V})$, the agent may test whether the hypothesized graph $\mathcal{G}$ is compatible with the available data. There exists a traditional literature known as *causal discovery* that attempts to perform the reverse process (Pearl, 2000; Spirtes et al., 2000; Petersen et al., 2006). In other words, from the marks readable from $P(\boldsymbol{V})$, the agent should infer what the compatible $\mathcal{G}$ that could have left these traces is. In practice, assumptions regarding the simplicity of these models (à la Occam's razor) are used to avoid situations described above, in which a saturated model would be preferred.

A growing, more recent literature is concerned with combining both observational and experimental distributions to learn a more restrictive equivalence class of causal diagrams (Kocaoglu et al., 2017, 2019; Wang et al., 2017; Agrawal et al., 2019; Mooij et al., 2020; Jaber et al., 2020). In fact, a richer set of constraints other than conditional independences emerge when we consider multiple distributions across different regimes (observational and interventional).

3. Elements of Causal Reinforcement Learning

In this section, we will introduce a unified framework that lets us view the decision-making problem through causal lenses and solve reinforcement learning tasks using causal inference tools. First, Sec. 3.1 formalizes the decision-making problem in the causal language by introducing a mathematical object called causal decision models (CDMs). Every CDM comprises of a structural causal model representing the underlying environment, a policy space encoding what the agent can control and observe during the intervention, and a reward function that gauges the agent's performance.

Armed with this new formalism, we can represent many canonical decision-making settings found in the literature within the semantic framework of SCMs. These include multi-armed bandits (MABs), Markov decision processes (MDPs), and dynamic treatment regimes (DTRs), among others. In practical scenarios, a detailed parametrization of the environment isn't always fully known, giving rise to reinforcement learning in SCMs. Sec. 3.2 introduces the concept of *causal reinforcement learning task*, and provides an initial catalog of tasks that will be studied through this section. Each task is delineated by the manner in which the agent interacts with the environment (regime),

any prior structural assumptions about that environment, and the specific policy space and reward function the agent seeks to optimize. Lastly, Sec. 3.3 delves into the importance of causal knowledge by examining policy learning within the conventional decision-making model of MDPs. Without explicitly acknowledging the learning regime and structural assumptions, we'll demonstrate that the underlying data-generating mechanisms can yield multiple MDPs, all compatible with the observed data, but with diverging implications for the optimization of decision-making. In simpler terms, the observed data typically doesn't fully dictate an optimal policy.

3.1 Causal Decision Models

We now formalize the policy optimization problem in SCMs based on the causal machinery introduced earlier in this section. The underlying environment will be represented as an SCM $\mathcal{M}^* = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P \rangle$. We study the problem of interacting on action variables $\boldsymbol{X} \subseteq \boldsymbol{V}$ to optimize some performance measures over reward signals $\boldsymbol{Y} \subseteq \boldsymbol{V}$ evaluated by $\mathcal{M}^*$. The agent determines values of actions $\boldsymbol{X}$ by performing intervention $\text{do}(\pi)$ following some policies π. The collection of all candidate policies π determining values of actions $\boldsymbol{X}$ defines a *policy space* Π. Formally,

Definition 8 (Policy Space) *For an SCM $\mathcal{M}^* = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P \rangle$, a policy space Π is a collection of policies π over actions $\boldsymbol{X} = \{X_1, \ldots, X_H\}$. Each policy π is a sequence of decision rules $(\pi_1 (X_1 \mid \boldsymbol{S}_1), \ldots, \pi_H (X_H \mid \boldsymbol{S}_H))$ such that for every $i = 1, \ldots, H$,*[16]

- *Action X_i is a non-descendent of $X_{i+1}, \ldots, X_H$, i.e., $X_i \in \boldsymbol{V} \setminus De(X_{i+1}, \ldots, X_H)$;*
- *States $\boldsymbol{S}_i$ are non-descendants of $X_i, \ldots, X_H$, i.e., $\boldsymbol{S}_i \subseteq \boldsymbol{V} \setminus De(X_i, \ldots, X_H)$.*

Henceforth, we will consistently denote such a policy space by $\Pi = \{\langle X_1, \boldsymbol{S}_1 \rangle, \ldots, \langle X_H, \boldsymbol{S}_H \rangle\}$. ■

Every policy $\pi \in \Pi$ is a sequence of decision rules $(\pi_1 (X_1 \mid \boldsymbol{S}_1), \ldots, \pi_H (X_H \mid \boldsymbol{S}_H))$. An agent following policy π selects values of actions $\boldsymbol{X}$ following a temporal ordering $X_1, \ldots, X_H$. At every step of intervention $i = 1, \ldots, H$, it performs the following

1. Observe some state variables $\boldsymbol{S}_i = \boldsymbol{s}_i$;
2. Select a value of action $x_i \sim \pi_i(X_i \mid \boldsymbol{S}_i = \boldsymbol{s}_i)$ following the decision rule π_i;
3. Perform an intervention $\text{do}(X_i \leftarrow x_i)$ following the selected action x_i.

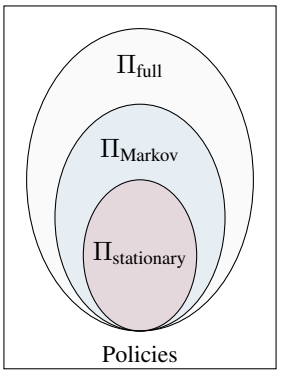


Figure 9: Policy spaces for a 2-stage DTR environment.

In words, a policy space Π defines the action space $\boldsymbol{X}$ that the agent could control after being deployed in the environment and the state space $\boldsymbol{S}_1, \ldots, \boldsymbol{S}_H$ that the agent could perceive at the time of intervention for every action $X_1, \ldots, X_H$. The subsequent examples illustrate the concept of policy space in dynamic treatment regimes (for short, DTRs), which is a class of sequential decision-making environments widely applied in healthcare and personalized medicine.

16. The policy space Π is also referred to a policy scope in (Lee and Bareinboim, 2020), which characterizes the agent's action and state scope - what it could interaction with and what it could observe at the time of interaction.

Example 12 (Dynamic Treatment Regimes (Murphy, 2003)) *In healthcare, a typical patient is often treated at multiple stages; the physician repeatedly adapts each treatment, tailoring it to the patient's time-varying, dynamic state. Dynamic treatment regimes provide an appealing framework for managing personalized medicine in the longitudinal setting.*

For instance, consider a DTR for managing alcohol-dependent patients, adapted from (Murphy et al., 2001a; Chakraborty and Moodie, 2013). The physician (i.e., the agent) has to decide the initial treatment X_1 and the secondary treatment X_2. Based on the condition of an alcohol-dependent patient (S_1), the physician may use behavioral therapy ($X_1 = 0$) or prescribe medication ($X_1 = 1$). The patient is then classified as a responder or a non-responder (S_2) based on the level of drinking within the next two months. The physician must then decide whether to continue the initial treatment ($X_2 = 0$) or switch to a more intensive plan combining medication and behavioral therapy ($X_2 = 1$). We are interested in the primary outcome Y that measures the patient's days of abstinence over 12 months after the treatment.

More formally, consider a DTR environment $\mathcal{M}^*_{\text{DTR}}$ that is a tuple given by

$$\mathcal{M}^*_{\text{DTR}} = \langle \boldsymbol{U} = \{U, U_1, \dots, U_4\}, \boldsymbol{V} = \{S_1, S_2, X_1, X_2, Y\}, \mathscr{F}_{\text{DTR}}, P(\boldsymbol{U})\rangle, \tag{77}$$

where the underlying causal mechanisms $\mathscr{F}_{\text{DTR}}$ are given by,

$$\mathscr{F}_{\text{DTR}} = \begin{cases} S_1 \leftarrow \mathbb{1}\{U_3 > 0\}, \\ X_1 \leftarrow \mathbb{1}\{3S_1 + \alpha_1 U + U_1 > 0\}, \\ S_2 \leftarrow \mathbb{1}\{0.1 + 0.1S_1 + 0.1X_1 + U_4 > 0\}, \\ X_2 \leftarrow \mathbb{1}\{3S_2 + \alpha_2 U + U_2 > 0\}, \\ Y \leftarrow \mathbb{1}\{3U - 3S_1 - 3X_1 - 3S_1X_1 + 3X_2 - 3S_2X_2 + 3X_1X_2 > 0\}, \end{cases} \tag{78}$$

Among quantities in the above equations, we set coefficients $\alpha_1 = \alpha_2 = 0$. However, for examples in subsequent sections, these coefficients will be non-zero.

The exogenous distribution $P(\boldsymbol{U})$ is defined such that for $i = 1, \dots, 4$, $U_i \sim \texttt{Logistic}(0, 1)$ is an independent variable drawn from a logistic distribution

$$P\left(U_i < u\right) = \frac{1}{1 + e^{-u}} \tag{79}$$

and $U \sim \texttt{Unif}(0, 1)$ is an independent variable drawn uniformly over a real interval $[0, 1]$. ■

It is possible to define different policy spaces in the DTR environment described above, depending on the choices of input states $\boldsymbol{S}_i$ for every action X_i, and parametric forms of decision rules π_i.

Example 13 (DTR, Policy Space) *For the 2-stage DTR environment described in Eq. 78. We consider a policy space $\Pi_{\textit{full}} = \{\langle X_1, \{S_1\}\rangle, \langle X_2, \{S_1, X_1, S_2\}\rangle\}$ satisfying the* perfect recall *(Koller and Friedman, 2009). That is, the agent determines every action based on all the past states and actions' history. More specifically, every adaptive treatment strategy (i.e., a policy) $\pi \in \Pi_{\textit{full}}$ is a pair of decision rules*

$$\pi = \left(\pi_1\left(X_1 \mid S_1\right), \pi_2\left(X_2 \mid S_1, X_1, S_2\right)\right), \tag{80}$$

where π_1 prescribes an initial treatment X_1 based on the patient's initial condition S_1; and π_2 decides whether to continue or switch the previous plan X_2 based on the patient's responses S_2 to the previous treatment, the initial treatment, and condition X_1, S_1.

One could also define a more restricted policy space $\Pi_{\textit{Markov}} = \{\langle X_1, \{S_1\}\rangle, \langle X_2, \{S_2\}\rangle\}$. *Every policy* $\pi \in \Pi_{\textit{Markov}}$ *is a pair of decision rules*

$$\pi = \left(\pi_1\left(X_1 \mid S_1\right), \pi_2\left(X_2 \mid S_2\right)\right), \tag{81}$$

where every π_i, $i = 1, 2$, *prescribes a treatment* X_i *based on the patient's condition* S_i *at the current stage. Such policies are also referred to as Markov policies in planning literature (Puterman, 1994). Note that for every Markov policy* $\pi' \in \Pi_{\textit{Markov}}$, *one could simulate it with a general policy* $\pi \in \Pi_{\textit{full}}$ *by setting* $\pi_2(X_2 \mid S_1, X_1, S_2) = \pi'_2(X_2 \mid S_2)$. *It follows that* $\Pi_{\textit{Markov}} \subset \Pi_{\textit{full}}$, *i.e., every Markov policy is contained in the general policy space* $\Pi_{\textit{full}}$.

Finally, we define a stationary policy space $\Pi_{\textit{stationary}} \subset \Pi_{\textit{Markov}}$. *Every policy* $\pi \in \Pi_{\textit{stationary}}$ *is a Markov policy satisfying an additional parametric constraint such that decision rule* π_i *remains invariant across all stages* $i = 1, 2$. *That is, for a stationary policy* $\pi = (\pi_1, \pi_2)$,

$$\pi_1\left(X_1 \mid S_1\right) = \pi_2\left(X_2 \mid S_2\right) \tag{82}$$

Since $\Pi_{\textit{Markov}} \subset \Pi_{\textit{full}}$, *we must have* $\Pi_{\textit{stationary}} \subset \Pi_{\textit{Markov}} \subset \Pi_{\textit{full}}$. *Fig. 9 shows a Venn diagram representing the relationship between* $\Pi_{\textit{stationary}}$, $\Pi_{\textit{Markov}}$, *and* $\Pi_{\textit{full}}$. *The outer rectangle represents all possible policies over actions* X_1 *and* X_2, *including over a singleton action* X_i, $i = 1, 2$. ■

The agent's performance is measured by a reward function that takes a set of reward signals in the environment as input.

Definition 9 (Reward Function) *For an SCM* $\mathcal{M}^* = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P\rangle$, *a reward function* $\mathcal{R}$ *is a function* $\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$ *mapping domains of a subset of endogenous variables* $\boldsymbol{Y} \subseteq \boldsymbol{V}$ *to a real value in* $\mathbb{R}$. *Moreover, the endogenous variables* $\boldsymbol{Y}$ *are called reward signals.* ■

Def. 9 covers most performance criteria in the decision-making literature. For instance, for a sequence of reward signals $\boldsymbol{Y} = \{Y_1, \ldots, Y_H\}$, the *cumulative reward* $\mathcal{R}_{\text{total}}(\boldsymbol{Y})$ is given by

$$\mathcal{R}_{\text{total}}(\boldsymbol{Y}) = \sum_{i=1}^{H} Y_i. \tag{83}$$

When the total number of reward signals $H \to \infty$, the above cumulative reward does not necessarily converge. In this case, a reasonable criterion is to consider the *average reward* given by

$$\mathcal{R}_{\text{average}}(\boldsymbol{Y}) = \frac{1}{H}\sum_{i=1}^{H} Y_i. \tag{84}$$

The above reward function is ensured to converge as the total number of reward signals $H \to \infty$. Alternatively, let a discount factor $\gamma \in (0, 1)$ and define the *discounted cumulative reward* as:

$$\mathcal{R}_{\text{discount}}(\boldsymbol{Y}) = \sum_{i=1}^{H} \gamma^{i-1} Y_i. \tag{85}$$

The discount factor can be interpreted in several ways; as an interest rate, the probability of living another step, or the mathematical trick for bounding the infinite sum.

Definition 10 (Causal Decision Model) *A causal decision model (CDM) is a tuple $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ where $\mathcal{M}^* = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P \rangle$ is an SCM, Π is a policy space over actions $\boldsymbol{X} \subseteq \boldsymbol{V}$, and $\mathcal{R}$ is a reward function over reward signals $\boldsymbol{Y} \subseteq \boldsymbol{V}$.* ■

Among elements in Def. 10, the SCM $\mathcal{M}^*$ represents the underlying environment; the policy space Π indicates the agent's capabilities after deployed in the environment, i.e., what it could control and observe at the time of interaction; the reward function $\mathcal{R}$ measures the performance of the agent, from the system designer's perspective. Formally, every CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ characterizes a planning/decision-making task (Bellman, 1966) that attempts to find an optimal strategy from the policy space Π dictating the agent's behaviors, provided with the complete parametrization of the underlying environment $\mathcal{M}^*$. An optimal policy π^* for a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ is a policy in space Π that maximizes the reward function $\mathcal{R}$ evaluated by the underlying SCM $\mathcal{M}^*$, i.e.,

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_\pi \left[\mathcal{R}\left(\boldsymbol{Y}\right); \mathcal{M}^* \right] \tag{86}$$

We will graphically represent every CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ using an augmented causal diagram $\mathcal{G}$ constructed from the environment $\mathcal{M}^*$ (Fig. 6); actions $\boldsymbol{X}$ and reward signals $\boldsymbol{Y}$ are highlighted in blue and red respectively; for every action, $X_i \in \boldsymbol{X}$, its input states $\boldsymbol{S}_i$ are highlighted in light blue. Fig. 10 shows the graphical representation for some canonical planning tasks.

A few observations are worth making at this point. First, the number of action variables $H = |\boldsymbol{X}|$ represents the horizon of the decision sequence.[17] When $H = 1$, the CDM corresponds to the single-stage decision models such as MABs (Robbins, 1952). On the other hand, when $H > 1$, the CDM defines a sequential decision-making problem, e.g., MDPs (Puterman, 1994), when the agent has to sequentially determine values of actions X_i, $i = 1, \ldots, H$, based on values of the observed states $\boldsymbol{S}_i$ at the time of the intervention.

Second, it is possible to define multiple CDMs in the same environment $\mathcal{M}^*$ by changing the policy space Π and the reward function $\mathcal{R}$, resulting in different optimal policies. In other words, the optimal policy π^* is defined with regard to the agent's capabilities to interact the environment after being deployed and how the system designer incentives its behaviors. Consequently, changing the forms of policy Π and reward $\mathcal{R}$ affects the optimal solution π^*.

Example 14 (CDMs in DTR) *Let $\mathcal{M}^*_{\text{DTR}}$ be a DTR environment compatible with Fig. 10c. The primary outcome Y is given by $Y \leftarrow X_2 \oplus S_1$; values of S_1 are uniformly drawn from the binary domain $\{0, 1\}$. Let Π_{full} and Π_{Markov} be policy spaces defined in Example 13.*

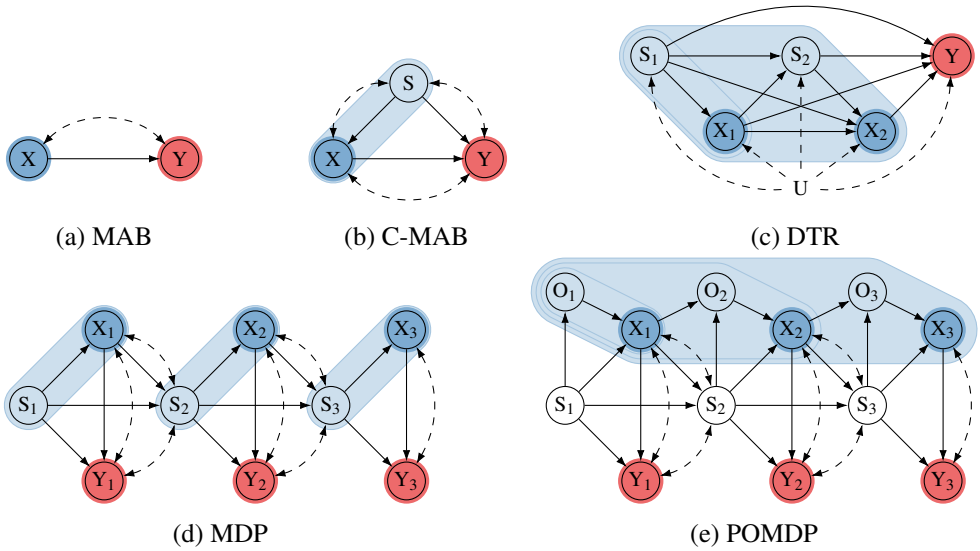
We could define CDMs $\langle \mathcal{M}^*_{\text{DTR}}, \Pi_{full}, Y \rangle$ and $\langle \mathcal{M}^*_{\text{DTR}}, \Pi_{Markov}, Y \rangle$, which represent two different planning tasks. The former searches for a general policy $\pi^*_{full} \in \Pi_{full}$ determining actions X_i based on the complete states and actions' history $S_1, \ldots, S_i, X_1, \ldots, X_{i-1}$; while the latter finds a Markov policy $\pi^*_{Markov} \in \Pi_{Markov}$ selecting action X_i based on the current state S_i.

First, for any Markov policy $\pi_{Markov} = (\pi_1, \pi_2)$ in Π_{Markov}, its expected reward is given by

$$\mathbb{E}_{\pi_{Markov}}[Y] = \sum_{x_2, s_1} \mathbb{E}[x_2 \oplus s_1] P(s_1) \sum_{s_2} \pi_2(x_2 \mid s_2) P(s_2) \tag{87}$$

$$= 0.5 \tag{88}$$

17. In episodic reinforcement learning, the agent collects data by interacting with the environment for repeated episodes $t = 1, \ldots, T$. The decision horizon H represents the total steps of actions $X_1, \ldots, X_H$ that the agent has to decide in every episode. We will further elaborate on the episodic learning for optimizing CDMs in Sec. 3.2.



(a) MAB (b) C-MAB (c) DTR

(d) MDP (e) POMDP

Figure 10: Causal diagrams for CDMs representing canonical decision-making models.

The last step holds since S_1 is uniformly drawn over the binary domain $\{0, 1\}$. Also, solving the CDM $\langle \mathcal{M}^*_{\text{DTR}}, \Pi_{full}, Y \rangle$ gives an optimal policy $\pi^*_{full} = (\pi^*_1, \pi^*_2)$ such that $\pi^*_2 \triangleq X_2 \leftarrow \neg S_1$. Evaluating the expected reward of π^*_{full} in $\mathcal{M}^*_{\text{DTR}}$ gives

$$\mathbb{E}_{\pi^*_{full}}[Y] = \mathbb{E}[S_1 \oplus \neg S_1] \tag{89}$$

Computing the above equation implies $\mathbb{E}_{\pi^*_{full}}[Y] = 1$ which outperforms the best possible Markov policy $\mathbb{E}_{\pi^*_{Markov}}[Y] = 0.5$. Moreover, suppose the sign of the reward function is changed to $\mathcal{R}(Y) \leftarrow -Y$. The same solution π^*_{full} is no longer optimal in a CDM $\langle \mathcal{M}^*_{\text{DTR}}, \Pi_{full}, -Y \rangle$ since it now minimizes the expected primary outcome $\mathbb{E}_{\pi}[-Y]$ instead. ■

More broadly, the formulation of CDMs permits one to represent canonical planning tasks (or equivalently, decision-making models) in the literature across disciplines, including RL and healthcare, when the detailed parametrization of the underlying environment $\mathcal{M}^*$ is known. These canonical tasks are graphically represented in Fig. 10, which we will briefly describe below.

- **Multi-Armed Bandit (Robbins, 1952).** Fig. 10a is induced by a MAB model $\langle \mathcal{M}^*_{\text{MAB}}, \Pi, Y \rangle$ consisting of an arm choice X and reward Y. Policy space $\Pi = \{\langle X, \emptyset \rangle\}$ defines a set of policies π that selects values of action X following a probability distribution $\pi(X)$. Please visit Example 15 for a detailed instance of a MAB model.

- **Contextual Bandit (Langford and Zhang, 2008a).** Fig. 10b is the graphical representation of a contextual bandit (C-MAB) model $\langle \mathcal{M}^*_{\text{CMAB}}, \Pi, Y \rangle$. Compared with MABs, a context variable S is now observed. The policy space $\Pi = \{\langle X, \{S\} \rangle\}$ consists of candidate policies $\pi(X|S)$ which selects values of action X based on the observed context S.

- **Dynamic Treatment Regime (Murphy et al., 2001a).** In a DTR model $\langle \mathcal{M}^*_{\text{DTR}}, \Pi, Y \rangle$, the policy space $\Pi = \{\langle X_i, \{S_1, \ldots, S_i, X_1, \ldots, X_{i-1}\} \rangle\}_{i=1}^{H}$ is a set of policies $\pi = (\pi_1, \ldots, \pi_H)$

consisting of a finite sequence of decision rules. For every i-th stage, the decision rule $\pi_i(X_i \mid S_1, \ldots, S_i, X_1, \ldots, X_{i-1})$ selects values of action X_i based on all the past action and states' history $S_1, \ldots, S_i, X_1, \ldots, X_{i-1}$. The goal is to maximize the primary outcome Y after intervening on all actions $X_1, \ldots, X_H$. Fig. 10c is the graphical representation of a 2-stage DTR model; a detailed instance is provided in Example 12.

- **Markov Decision Process (Bellman, 1957; Puterman, 1994).** Consider a Markov decision process (MDP) model $\langle \mathcal{M}^*_{\text{MDP}}, \Pi, \mathcal{R} \rangle$. Environment $\mathcal{M}^*$ consists of a set of states $\boldsymbol{S} = \{S_i\}_{i=1}^{\infty}$, a set of actions $\boldsymbol{X} = \{X_i\}_{i=1}^{\infty}$, and a set of reward signals $\boldsymbol{Y} = \{Y_i\}_{i=1}^{\infty}$. The policy space $\Pi = \{\langle X_i, S_i \rangle\}_{i=1}^{\infty}$ consists of a set of decision rules $\pi = (\pi_i(X_i \mid S_i))_{i=1}^{\infty}$. The reward function $\mathcal{R}$ can be described as the average reward $\mathcal{R}_{\text{average}}(\boldsymbol{Y})$ in Eq. 84 or the discounted reward $\mathcal{R}_{\text{discount}}(\boldsymbol{Y})$ in Eq. 85. In the discounted case, rewards obtained later are discounted more than rewards obtained earlier. If the discounted factor $\gamma = 0$, the agent is said to be myopic, i.e., it is only concerned about immediate rewards. Fig. 10d represents the causal diagram of an MDP model spanning over steps $i = 1, 2, 3$. See Example 16 for a detailed instance of a policy planning task in an MDP environment.

- **Partially Observable MDP (Åström, 1965).** A partially observable MDP (POMDP) is a generalization of MDP in which system dynamics are determined by an MDP environment, but the agent could not directly utilize the underlying state $\boldsymbol{S} = \{S_i\}_{i=1}^{\infty}$ as input to determine its actions $\boldsymbol{X} = \{X_i\}_{i=1}^{\infty}$. Instead, it only receives a set of observation variables $\boldsymbol{O} = \{O_i\}_{i=1}^{\infty}$ depending on the underlying states. Fig. 10e shows a POMDP model $\langle \mathcal{M}^*_{\text{POMDP}}, \Pi, \mathcal{R} \rangle$ spanning over steps $i = 1, 2, 3$. The policy space $\Pi = \{\langle X_i, \{O_1, \ldots, O_i, X_1, \ldots, X_{i-1}\} \rangle\}_{i=1}^{\infty}$ defines a set of non-Markov policies $\pi = (\pi_i(X_i \mid O_1, \ldots, O_i, X_1, \ldots, X_{i-1}))_{i=1}^{\infty}$. For every i-th stage of intervention, an agent following a non-Markov policy $\pi \in \Pi$ selects an action $x_i \sim \pi_i(X_i \mid O_1, \ldots, O_i, X_1, \ldots, X_{i-1})$ based on all the past observations and actions.

When the policy space Π and the reward function $\mathcal{R}$ are well-specified, and detailed parameters of the underlying environment $\mathcal{M}^*$ are provided, there exist efficient algorithms in the planning literature to solve for an optimal policy in a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ (Bellman, 1957; Puterman, 1994; Shachter, 1986). For instance, for an MDP model graphically described in Fig. 10d, one could obtain an optimal policy using standard dynamic programming algorithms (Bellman, 1957; Puterman, 1994). The same planning procedure applies to a DTR model (Murphy, 2003; Murphy et al., 2001a; Murphy, 2005b), e.g., Fig. 10c. Due to the latent nature of underlying states, planning in POMDP models (e.g., 10e) is more computationally challenging, and requires the planning algorithm to maintain memory and possibly reason about beliefs over the states. A variety of heuristics for approximate planning in POMDPs have been proposed (Jaakkola et al., 1994; Hansen, 1998; Hauskrecht, 2000). Optimizing policies in a general CDM has been studied under the rubrics of influence diagrams; several algorithms and approximate procedures have been proposed, including (Shachter, 1986; Koller and Milch, 2003; Lauritzen and Nilsson, 2001).

The following examples illustrate CDMs in some canonical decision-making settings in the literature, together with the planning procedure for computing the optimal policy.

Example 15 (MAB Planning) *Consider a CDM described by the tuple*

$$\langle \mathcal{M}^* = \mathcal{M}^*_{\text{MAB}}, \Pi = \{\langle X, \emptyset \rangle\}, \mathcal{R}(Y) = Y \rangle, \tag{90}$$

where $\mathcal{M}^*_{\text{MAB}}$ is the MAB environment described in Example 1. Fig. 10a shows the graphical representation of this CDM where action X and reward Y are highlighted in blue and red respectively.

Every policy $\pi(X) \in \Pi$ is a probability distribution over domains of action $\mathscr{D}(X) = \{0, 1\}$. Evaluating the expected reward Y in submodel $\mathcal{M}^*_{\text{MAB}_\pi}$ gives

$$\mathbb{E}_\pi[Y] = \sum_x \sum_u \mathbb{E}[Y \mid x, u]\, P(u)\pi(x) \tag{91}$$

$$= \mathbb{E}_{X \leftarrow 0}[Y]\, \pi(X = 0) + \mathbb{E}_{X \leftarrow 1}[Y]\, \pi(X = 1) \tag{92}$$

The last step follows from marginalizing over the exogenous variable U. Note that evaluation of expected rewards $\mathbb{E}_x[Y]$ of atomic interventions $do(X \leftarrow x)$ is provided in Example 7. Replacing interventional queries $\mathbb{E}_x[Y]$ in the above equation gives

$$\mathbb{E}_\pi[Y] = 0.4\pi(X = 0) + (0.4 - \Delta)\pi(X = 1) \tag{93}$$

$$= 0.4 - \Delta\pi(X = 1) \tag{94}$$

The last step follows from $\sum_x \pi(x) = 1$. Since the coefficient $\Delta > 0$, the reward function $\mathcal{R}(Y) = Y$ evaluated in $\mathcal{M}^*_{\text{MAB}}$ is maximized when probability $\pi(X = 1) = 0$, That is, the optimal policy is determinsitic $\pi^* : X \leftarrow x^*$ with the optimal arm choice $x^* = 0$. ■

Example 16 (MDP Planning, Dynamic Programming) *Consider the MDP environment $\mathcal{M}^*_{\text{MDP}}$ given by Eq. 5. We are interested in optimizing the MDP model described by a CDM given by*

$$\langle \mathcal{M}^* = \mathcal{M}^*_{\text{MDP}}, \Pi = \{\langle X_i, \{S_i\}\rangle\}_{i=1}^{\infty}, \mathcal{R}_{\textit{discount}}(\boldsymbol{Y})\rangle \tag{95}$$

where $\mathcal{R}_{\textit{discount}}(\boldsymbol{Y})$ is the discounted reward function given by Eq. 85 with $\gamma = 0.9$. We will focus on stationary policies $\pi = (\pi_i(X_i \mid S_i))_{i=1}^{\infty}$ such that decision rules $\pi_1 = \pi_2 = \dots$ remain invariant across decision horizons $i = 1, 2, \dots$.[18]

Let $\mathscr{D}(S)$ and $\mathscr{D}(X)$ denote the domain of state S_i and action X_i at every stage i, respectively. For any stationary policy $\pi \in \Pi$, a state-action value function $Q_\pi : \mathscr{D}(S) \times \mathscr{D}(X) \to \mathbb{R}$ (also called a Q-function) is defined as the expected cumulative reward following policy π given the starting state s and initial action x, i.e.,

$$Q_\pi(s, x) = \mathbb{E}_\pi\left[\sum_{j=0}^{\infty} \gamma^j Y_{i+j} \mid S_i = s, X_i = x\right] \tag{96}$$

Since the structural functions f_{S_i} and f_{Y_i} remains invariant across decision horizon $i = 1, 2, \dots$, for any policy π, any state s, and any action x, the above expression can be recursively defined in terms of a so-called Bellman Equation *(Bellman, 1966):*

$$Q_\pi(s, x) = \mathbb{E}_\pi\left[Y_i + \gamma Y_{i+1} + \gamma^2 Y_{i+2} + \cdots \mid S_i = s, X_i = x\right] \tag{97}$$

$$= \mathbb{E}_\pi[Y_i + \gamma Q_\pi(S_{i+1}, X_{i+1}) \mid S_i = s, X_i = x] \tag{98}$$

18. Indeed, it has been shown that there always exists a stationary policy that could optimize the cumulative reward in an MDP model (Filar and Vrieze, 2012). It thus suffices to focus on stationary policies.

S	X	Q_*	S	X	Q_*
0	0	**8.2**	1	0	7.56
0	1	7.56	1	1	**8.2**

Table 2: Optimal Q-function $Q_*(s, x)$ evaluated in the MDP model of Example 16.

The last step follows from the recursive definition of the value function $Q_\pi(s, x)$ and the Markov property in the interventional distribution; see Example 9 for details. The above equation could be further written as, by expanding on next state S_{i+1} and action X_{i+1},

$$Q_\pi(s,x) = \mathbb{E}_\pi\left[Y_i \mid S_i = s, X_i = x\right] + \gamma \mathbb{E}_\pi\left[Q_\pi(S_{i+1}, X_{i+1}) \mid S_i = s, X_i = x\right] \tag{99}$$

$$= \mathbb{E}_\pi\left[Y_i \mid S_i = s, X_i = x\right] \tag{100}$$

$$+ \gamma \sum_{s'} P_\pi\left(S_{i+1} = s' \mid S_i = s, X_i \leftarrow x\right) \sum_{x'} \pi_{i+1}(x' \mid s') Q_\pi(s', x') \tag{101}$$

Since the transition distribution and the conditional reward remain invariant across atomic and policy interventions (Eqs. 60 and 63), the above equation could be further written as

$$Q_\pi(s,x) = \mathcal{R}_{exp}(s,x) + \gamma \sum_{s'} \mathcal{T}_{exp}(s,x,s') \sum_{x'} \pi_{i+1}(x' \mid s') Q_\pi(s', x') \tag{102}$$

where the transition probability $\mathcal{T}_{exp}$ and the reward function $\mathcal{R}_{exp}$ are given by

$$\mathcal{T}_{exp}(s,x,s') = P_{X_i \leftarrow x}\left(S_{i+1} = s' \mid S_i = s\right) \tag{103}$$

$$\mathcal{R}_{exp}(s,x) = \mathbb{E}_{X_i \leftarrow x}\left[Y_i \mid S_i = s\right] \tag{104}$$

Detailed parametrizations of interventional quantities $\mathcal{T}$ and $\mathcal{R}$ are provided in Fig. 6.

An optimal policy π^* is such that $Q_{\pi^*}(s,x) \geq Q_\pi$ for all state-action pair s, x and all policies π. Optimizing Q-function leads to an expression called the Bellman optimality equation*, i.e,*

$$Q_*(s,x) = \mathcal{R}_{exp}(s,x) + \gamma \sum_{s'} \mathcal{T}_{exp}(s,x,s') \max_{x'} Q_*(s', x') \tag{105}$$

The optimal policy π^* is given by, for every stage $i = 1, 2, \ldots,$

$$\pi_i^*(S_i = s) = \arg\max_x Q_*(s,x) \tag{106}$$

for any state $s \in \mathscr{D}(S)$. We can compute the optimal Q-function evaluated in $\mathcal{M}^*_{\text{MDP}}$ using value iteration (Sutton and Barto, 1998). Detailed parametrizations are provided in Table 2. Complete computations are provided in Appendix B. The optimal policy $\pi^* = (\pi_i^*(X_i \mid S_i))_{i=1}^\infty$ is given by $\pi_i^* \triangleq X_i \leftarrow S_i$, for every $i = 1, 2, \ldots$. Evaluating its expected return gives $\mathbb{E}_{\pi^*}\left[\sum_{i=1}^\infty \gamma^{i-1} Y_i\right] = 8.2$; detailed derivation steps are provided in Example 10. ■

Example 17 (DTR Planning, Dynamic Programming) *Consider the DTR environment $\mathcal{M}^*_{\text{DTR}}$ in Eq. 78. Our goal is to maximize the patient's days of abstinence Y over 12 months after the treatment. This decision-making problem is described by a CDM given by*

$$\langle \mathcal{M}^* = \mathcal{M}^*_{\text{DTR}}, \Pi = \{\langle X_1, \{S_1\}\rangle, \langle X_2, \{S_1, X_1, S_2\}\rangle\}, \mathcal{R}(Y) = Y \rangle \tag{107}$$

S_1	X_1	$Q_*^{(1)}$	S_1	X_1	$Q_*^{(1)}$
0	0	**0.8799**	1	0	**0.4716**
0	1	0.8749	1	1	0.1003

(a) $Q_*^{(1)}(s_1, x_1)$

S_1	X_1	S_2	X_2	$Q_*^{(2)}$	S_1	X_1	S_2	X_2	$Q_*^{(2)}$
0	0	0	0	0.7851	1	0	0	0	0.2149
0	0	0	1	**0.9846**	1	0	0	1	**0.7851**
0	0	1	0	0.7851	1	0	1	0	0.2149
0	0	1	1	**0.7851**	1	0	1	1	**0.2149**
0	1	0	0	0.2149	1	1	0	0	0.0008
0	1	0	1	**0.9846**	1	1	0	1	**0.2149**
0	1	1	0	0.2149	1	1	1	0	0.0008
0	1	1	1	**0.7851**	1	1	1	1	**0.0154**

(b) $Q_*^{(2)}(s_1, x_1, s_2, x_2)$

Table 3: Evaluation of optimal Q-functions evaluated in the DTR system of Example 12.

Fig. 10c describes a causal diagram of $\mathcal{M}^*_{\text{DTR}}$ where actions X_1, X_2 are highlighted in blue, primary outcome Y in red, and input covariates $\{S_1\}$ and $\{S_1, X_1, S_2\}$ in light blue.

For every policy $\pi \in \Pi$, evaluating recovery rate Y in submodel $\mathcal{M}^*_{\text{DTR}_\pi}$ gives

$$\mathbb{E}_\pi[Y] = \sum_{s_1,x_1,s_2,x_2} \pi_2(x_2 \mid s_1, x_1, s_2)\pi_1(x_1 \mid s_1)\Bigg(\sum_u \mathbb{E}[Y \mid s_1, x_1, s_2, x_2, u] \\ P(s_2 \mid s_1, x_1, u)\, P(s_1 \mid u)P(u)\Bigg) \quad (108)$$

Since π_1, π_2 are not functions of the exogenous variable U, summing over domain of U we obtain

$$\mathbb{E}_\pi[Y] = \sum_{s_1,x_1,s_2,x_2} \mathbb{E}_{x_1,x_2}[Y \mid s_1, s_2]\, \pi_2(x_2 \mid s_1, x_1, s_2)P_{x_1}(s_2 \mid s_1)\, \pi_1(x_1 \mid s_1)P(s_1) \quad (109)$$

As shown in (Murphy et al., 2001a), the optimal policy π^* is deterministic, and satisfies the Bellman equation (Bellman, 1957)

$$\pi_1^*(s_1) = \arg\max_{x_1} Q_*^{(1)}(s_1, x_1) \quad (110)$$

$$\pi_2^*(s_1, x_1, s_2) = \arg\max_{x_2} Q_*^{(2)}(s_1, s_2, x_1, x_2) \quad (111)$$

The optimal Q-function is

$$Q_*^{(1)}(s_1, x_1) = \sum_{s_2} \max_{x_2} Q_*^{(2)}(s_1, s_2, x_1, x_2)\, P_{x_1}(s_2 \mid s_1) \quad (112)$$

$$Q_*^{(2)}(s_1, s_2, x_1, x_2) = \mathbb{E}_{x_1,x_2}[Y \mid s_1, s_2] \quad (113)$$

We compute the parametrization of Q-functions evaluated in $\mathcal{M}^*$ and provide them in Table 3a. See Appendix B for complete computation. The optimal actions $\pi_1^*(s_1)$, $\pi_2^*(s_1, x_1, s_2)$ given every state-action's history is highlighted. Solving for an optimal policy gives $\pi_1^* \triangleq X_1 \leftarrow 0$ and $\pi_2^* \triangleq X_2 \leftarrow 1$. In words, in order to maximize the patient's days of abstinence, the physician should start with behavioral therapy and follow up with an intensive treatment combining both behavioral therapy and medication. The expected reward of this policy π^* is computable as:

$$\mathbb{E}_{\pi^*}[Y] = \sum_{s_1} Q_*^{(1)}(s_1, X_1 = 0)P(s_1) \tag{114}$$

$$= Q_*^{(1)}(S_1 = 0, X_1 = 0)P(S_1 = 0) + Q_*^{(1)}(S_1 = 1, X_1 = 0)P(S_1 = 1) \tag{115}$$

Evaluating the above equation gives the optimal expected reward $\mathbb{E}_{\pi^*}[Y] = 0.6758$. ■

3.2 Causal Reinforcement Learning Tasks

The causal decision model described so far assumes the full knowledge of the underlying environment. However, in many real-world practical applications, the detailed parametrization of the environment is very rarely known, which means that standard planning algorithms are not immediately applicable. In order for the agent to optimize the performance of the underlying system, a learning process must take place, leading to the learning paradigm of *causal reinforcement learning*.

To make the argument more precise, we start the discussion of a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$. Recall that it defines a planning task of finding an optimal policy in space Π that maximizes the reward function $\mathcal{R}$ evaluated in the environment $\mathcal{M}^*$. A CRL agent $\mathbb{C}$ is assumed to have access to the policy space Π and the performance measurement $\mathcal{R}$.[19] However, the underlying environment $\mathcal{M}^*$ is not fully known. Instead, the agent $\mathbb{C}$ only has access to some structural assumptions $\mathcal{A}$ encoding qualitative knowledge about the environment $\mathcal{M}^*$, and a learning regime $\mathcal{L}$ dictating how it interacts with the environment $\mathcal{M}^*$ to collect data. This partial knowledge constitutes new dimensions for the task formulation of optimal decision-making under uncertainty, which we will briefly discuss below.

Learning Regimes ($\mathcal{L}$). Following the discussion of the PCH, each CRL agent may be able to interact with $\mathcal{M}^*$ in different ways, including through passive observations (i.e., $\mathcal{L} = \text{see}$) or by active interventions ($\mathcal{L} = \text{do}$). These learning regimes model distinct types of interactions of the agent with the environment. The former corresponds to off-policy reinforcement learning tasks (Li et al., 2011, 2014); while the latter corresponds to online reinforcement learning tasks (Auer et al., 2002b). More specifically, an agent passively observing the environment does not actively determine actions. Instead, it receives observational data $\mathcal{D} \sim P(\boldsymbol{V})$ summarizing trajectories of another agent (e.g., a human demonstrator) already operating in the environment, following a behavioral policy. On the other hand, an agent may actively control actions $\boldsymbol{X}$ by performing interventions $\text{do}(\pi)$, following some policies π, and receiving experimental data $\mathcal{D} \sim P_\pi(\boldsymbol{V})$. We will investigate these reinforcement learning algorithms in depth later on in this paper.

Structural Assumptions ($\mathcal{A}$). Structural assumptions specify a hypothesis class of possible environmental models that the agent is operating with. One common way of specifying assumptions

19. The policy space Π and the reward function $\mathcal{R}$ are provided in most of the learning tasks considered in this paper. However, there exist practical applications where the reward function $\mathcal{R}$ is not fully known. In this case, the agent has to "guess" a surrogate reward function from a hypothesis class $\mathbb{R}$ and then compute an optimal policy estimate with it. This setting is studied under the rubric of causal imitation learning in Sec. 8.

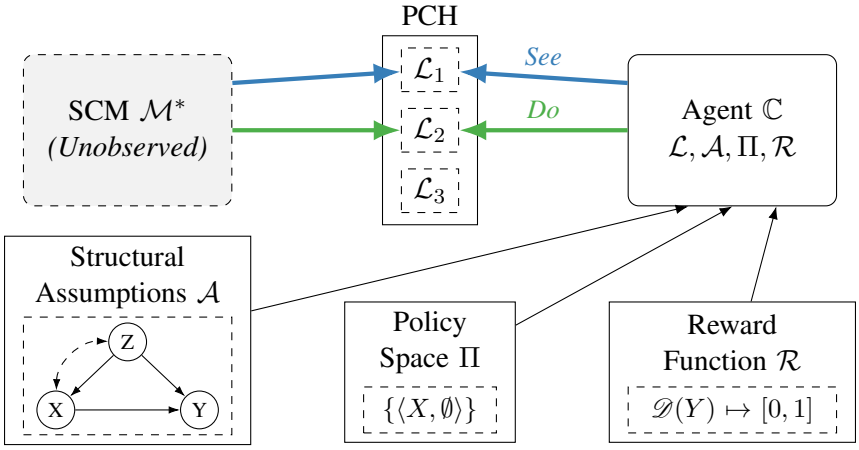


Figure 11: Graphical representation of a causal reinforcement learning task

about the SCM $\mathcal{M}^*$ is through a causal diagram $\mathcal{G}$ (Def. 6). The hypothesis class $\mathbb{M}$ is thus defined as the family of SCMs $\mathcal{M}$ compatible with the diagram $\mathcal{G}$, i.e., $\mathcal{G}(\mathcal{M}) = \mathcal{G}$. For instance, the causal diagram in Fig. 10d specifies a family of MDP environments consisting of states S_i, actions X_i, and reward signals Y_i, for $i = 1, 2, \ldots$; the underlying transition and reward distributions remain un-specified. Other assumptions include constraints over the features of the behavior policy (Pearl and Robins, 1995), or equivalence classes of causal diagrams (Zhang, 2008), to cite a few.

Consider again a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$, where the model of the environment $\mathcal{M}^*$ is not fully revealed to the agent. Replacing $\mathcal{M}^*$ with the learning regime $\mathcal{L}$ and structural assumptions $\mathcal{A}$ leads to a new signature $\langle \mathcal{L}, \mathcal{A}, \Pi, \mathcal{R} \rangle$ which characterizes a *causal reinforcement learning task*. Formally,

Definition 11 (Causal Reinforcement Learning Task) *For an SCM $\mathcal{M}^* = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P \rangle$, a CRL task $\mathcal{T}$ in the environment $\mathcal{M}^*$ is a 4-tuple $\langle \mathcal{L}, \mathcal{A}, \Pi, \mathcal{R} \rangle$, where*

1. *$\mathcal{L}$ is a learning regime of an agent's interaction with the SCM $\mathcal{M}^*$, possibly* see *or* do*;*
2. *$\mathcal{A}$ is a set of structural assumptions about the SCM $\mathcal{M}^*$;*
3. *Π is a policy space over actions $\boldsymbol{X}$;*
4. *$\mathcal{R}$ is a reward function over reward signals $\boldsymbol{Y}$.* ■

Formally, every CRL task $\langle \mathcal{L}, \mathcal{A}, \Pi, \mathcal{R} \rangle$ describes an optimal decision-making problem under uncertainties about the underlying environment. Provided with input $\langle \mathcal{L}, \mathcal{A}, \Pi, \mathcal{R} \rangle$, the CRL agent attempts to estimate an optimal policy $\pi^* \in \Pi$ defined in Eq. 86 maximizing the reward $\mathcal{R}$ evaluated in the unknown environment $\mathcal{M}^*$. Since $\mathcal{M}^*$ is not fully observed, we substitute it with the learning regime $\mathcal{L}$ and structural assumptions $\mathcal{A}$ about the environment, depending on the specific task. The goal of the agent is then to find a policy π^* such that

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R}(\boldsymbol{Y}) \,\middle|\, \mathcal{A}, \mathcal{L} \right] \tag{116}$$

Compared to the optimization given by Eq. 86, we move the unobserved SCM that evaluates the agent as a superscript and leave what the agent has access to as part of the conditioning set.

Algorithm 1 Causal Reinforcement Learning Agent $\mathbb{C}$

Require: CRL task $\mathcal{T} = \langle \mathcal{L}, \mathcal{A}, \Pi, \mathcal{R} \rangle$
Ensure: a policy estimate $\hat{\pi} \in \Pi$ optimizing a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$.
1: Let data $\mathcal{D} = \{\}$.
2: **for all** every episode $t = 1, \ldots, T$ **do**
3: Interact with the SCM $\mathcal{M}^*$ following regime $\mathcal{L}$ and receive samples $\boldsymbol{V}^{(t)} \sim P_{\mathcal{L}}\left(\boldsymbol{V}; \mathcal{M}^*\right)$.
4: Update data $\mathcal{D} = \mathcal{D} \cup \left\{\boldsymbol{V}^{(t)}\right\}$.
5: **end for**
6: **return** an empirical estimate $\hat{\pi} \in \Pi$ of the optimal policy π^* from data $\mathcal{D}$ and assumptions $\mathcal{A}$.

For instance, Fig. 11 shows a graphical instance of a CRL task $\mathcal{T}$ in an unknown bandit environment $\mathcal{M}^*$ consisting of an arm choice X, reward Y, and a covariate Z. The goal is to find an optimal policy $X \leftarrow x^*$ in the policy space $\Pi = \{X, \emptyset\}$ maximizing the expected reward $\mathbb{E}_x\left[\mathcal{R}(Y)\right]$. The agent could passively observe the environment ($\mathcal{L} =$ See) and receives observational data drawn from $P(X, Y, Z)$; it could also actively intervene on the arm ($\mathcal{L} =$ Do) and receive interventional data drawn from $P_x\left(Y, Z\right)$. The causal diagram $\mathcal{A} = \mathcal{G}_{\text{backdoor}}$ (bottom left) encodes structural knowledge that the agent has about the environment: there is no spurious correlation between Z and Y and between X and Y.

Alg. 1 provides pseudo-code describing the general learning strategy of a CRL agent to solve a task $\langle \mathcal{L}, \mathcal{A}, \Pi, \mathcal{R} \rangle$. We follow the episodic reinforcement learning setting (Sutton and Barto, 1998) where the agent interacts with the environment $\mathcal{M}^*$ for repeated episodes $t = 1, \ldots, T$. For each episode t, the agent interacts with the environment following the learning regime $\mathcal{L}$ and receives sample $\boldsymbol{V}^{(t)}$, consisting of realized actions $\boldsymbol{X}^{(t)}$, observed states $\boldsymbol{S}^{(t)}$, reward signals $\boldsymbol{Y}^{(t)}$, and other endogenous variables. For the observational regime $\mathcal{L} =$ *see*, the CRL agent passively observes another agent, currently deployed, to determine values of every action $X_i \in \boldsymbol{X}$ following a behavioral policy $f_{\boldsymbol{X}}$. For the interventional regime $\mathcal{L} =$ *do*, the CRL agent actively intervenes on every action $X_i \in \boldsymbol{X}$ following a policy π, and receives subsequent states $\boldsymbol{S}$ and rewards $\boldsymbol{Y}$. Finally, the agent $\mathbb{C}$ computes an empirical estimate of an optimal policy π^* in the policy space Π from the combination of the collected data $\mathcal{D}$ and structural assumptions $\mathcal{A}$.

The formulation of the task signature and the CRL agent summarizes existing policy learning problems and learning strategies in the reinforcement learning and causal inference literature. The following examples illustrate CRL tasks in single-stage decision-making settings.

Example 18 (Off-Policy Learning (Sutton and Barto, 1998), MAB) *Consider first a MAB model $\langle \mathcal{M}^*, \{\langle X, \emptyset \rangle\}, Y \rangle$ graphically described in Fig. 10a. A CRL agent $\mathbb{C}$ aims to learn an arm*

$$x^* = \arg\max_x \mathbb{E}_x\left[Y\right] \tag{117}$$

with the maximal expected reward. The detailed parametrization of SCM $\mathcal{M}^*$ is unknown. Instead, the agent could only passively observe the environment and receive the observational distribution $P(X, Y)$ evaluated in $\mathcal{M}^*$. This leads to an off-policy learning *task described through the following signature:*

$$\mathcal{T}_{\mathit{off}} = \langle \mathcal{L} = \mathit{see}, \mathcal{A} = \mathit{NUC}, \Pi = \{\langle X, \emptyset \rangle\}, \mathcal{R} = Y \rangle, \tag{118}$$

NUC stands for the assumption of "No Unmeasured Confounder": there is no unobserved confounder affecting the action X and the reward Y simultaneously. All the observed correlations between X and Y are fully explained by the causal relationships among them, which implies that

$$P_x(Y) = P(Y \mid X = x) \tag{119}$$

Therefore, the agent could evaluate the expected reward of every arm x from the observational data $P(X, Y)$[20], and find optimal an optimal arm x^* with the maximal empirical reward estimates. ■

Example 19 (Online Learning (Sutton and Barto, 1998), MAB) *We will continue with the previous example of the MAB model $\langle \mathcal{M}^*, \{\langle X, \emptyset\rangle\}, Y\rangle$. Suppose the CRL can now actively intervene in the underlying environment $\mathcal{M}^*$. This leads to an* online learning *task described by a signature*

$$\mathcal{T}_{on} = \langle \mathcal{L} = do, \mathcal{A} = \emptyset, \Pi = \{\langle X, \emptyset\rangle\}, \mathcal{R} = Y\rangle, \tag{120}$$

In order to evaluate every arm x, the CRL agent performs an intervention $do(x)$ in SCM $\mathcal{M}^*$ and receives subsequent reward signals drawn from $P_x(Y)$. The expected reward $\mathbb{E}_x[Y]$ is thus estimable from the experimental data by computing the empirical means. ■

Example 20 (Causal Identification (Pearl, 2000), MAB) *Consider the off-policy learning task of Eq. 118 again. Suppose now that the NUC assumption no longer holds. Instead, a previously unobserved covariate Z is now revealed; the CRL agent has access to a more detailed causal diagram $\mathcal{G}_{backdoor}$ (Fig. 11, bottom left) describing the underlying environment $\mathcal{M}^*$. Replacing the NUC with structural assumptions encoded in diagram $\mathcal{G}_{backdoor}$ leads to a* causal identification *task*

$$\mathcal{T}_{id} = \langle \mathcal{L} = see, \mathcal{A} = \mathcal{G}_{backdoor}, \Pi = \{\langle X, \emptyset\rangle\}, \mathcal{R} = Y\rangle, \tag{121}$$

Like off-policy learning, the CRL agent observes the environment and receives the observational distribution $P(X, Y, Z)$ evaluated in $\mathcal{M}^*$. Provided with the causal diagram $\mathcal{G}_{backdoor}$, applying the backdoor adjustment formula (Pearl, 2000, Ch. 3.3) implies

$$P_x(y) = \sum_z P(y \mid z, x) P(x) \tag{122}$$

That is, the expected reward of every arm x is computable from the observational data $P(X, Y, Z)$. Optimizing the expected reward over the action domain X leads to an optimal arm. ■

Broadly speaking, the formalization of the CRL dimensions and the corresponding tasks, semantically defined through structural causal models, allows us to describe most of the popular learning settings studied in the literature. It also enables us to explore and study novel CRL learning tasks beyond the current literature and that arises naturally in real-world applications. We first summarize the more traditional RL-CI tasks with their corresponding signatures in Table 4. Specifically, we will discuss in Sec. 4, the policy learning methods for traditional CI and RL tasks through the language of CRL. In practice, this encompasses tasks such as off-policy and online learning and causal identification. All of these focus on optimizing policies within an experimental policy space Π_{EXP} (Def. 8). These tasks can be viewed as variations of the first two dimensions described in

20. We assume that the number of the observational data is sufficient, and joint distribution $P(X, Y)$ is recovered.

		Signature				
	Task	**Learning Regime** ($\mathcal{L}$)	**Structural Assumptions** ($\mathcal{A}$)	**Policy Space** (Π)	**Reward Function** ($\mathcal{R}$)	**Section**
1	Off-policy Learning	See	NUC	Π_{EXP}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	4.1
2	Online Learning	Do	-	Π_{EXP}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	4.2
3	Causal Identification	See	DAG $\mathcal{G}$	Π_{EXP}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	4.3

Table 4: Summary of causal reinforcement learning tasks investigated in this paper, in terms of their signatures and sections. We highlight in gray the most distinct feature introduced by the task.

the table, namely, the interactive regime and the structural assumptions, while the other dimensions remain constant. Even though these tasks are prevalent in current literature, certain critical conditions weren't formally understood prior to our new formalization. For instance, determining the validity of an off-policy method (e.g., inverse propensity weighting and dynamic programming), specifically, whether it can find a policy consistent with optimal one given by the underlying SCM. Analyzing these classical tasks through CRL perspective will be instrumental in illuminating other foundational issues and illustrating how causal and RL formalisms intersect.

3.3 Comparison with Markov Decision Processes

The CRL tasks described so far (Def. 11) assume that the agent has access to either the learning regime $\mathcal{L}$ under which the data are collected, or structural assumptions $\mathcal{A}$ encoding causal invariances about the environment. In this section, we will demonstrate that such causal knowledge is generally indispensable (necessary) for learning an optimal policy in an unknown SCM. Our discussion focuses on standard MDPs, which is a class of sequential decision-making models widely used in practice.

Definition 12 (Standard MDP (Puterman, 1994)) *A Markov decision process is tuple $\langle \mathscr{D}(S), \mathscr{D}(X), \mathcal{T}, \mathcal{R} \rangle$ where*

1. *$\mathscr{D}(S)$ is a set of states called the state space;*
2. *$\mathscr{D}(X)$ is a set of actions called the action space;*
3. *$\mathcal{T}(s, x, s') \in [0, 1]$ is a transition probability that action $X_i = x$ in state $S_i = s$ at stage i will lead to state $S_{i+1} = s'$ at stage $i + 1$;*
4. *$\mathcal{R}(s, x)$ is the immediate reward received in state $S_i = s$ due to action $X_i = x$ at stage i.*

A policy $\pi(x|s)$ in a standard MDP is a function mapping from the state space $\mathcal{D}(S)$ to a probability distribution over the action space $\mathcal{D}(X)$. For any indices $i < j \in \mathbb{N}$, let $\bar{\boldsymbol{V}}_{i:j}$ denote a sequence $\{V_i, V_{i+1}, \ldots, V_j\}$. Given a policy π and a distribution over the initial state $P(S_1)$, every standard

S	X	Q_*	S	X	Q_*
0	0	9	1	0	**9**
0	1	**9**	1	1	9

Table 5: Optimal Q-function $Q_*(s, x)$ evaluated in the MDP model of Example 21.

MDP model defines a joint distribution over states $\bar{\boldsymbol{S}}_{1:H}$, actions $\bar{\boldsymbol{X}}_{1:H}$, and rewards $\bar{\boldsymbol{Y}}_{1:H}$ up to decision horizon H, i.e.,

$$P_\pi(\bar{\boldsymbol{s}}_{1:H}, \bar{\boldsymbol{x}}_{1:H}, \bar{\boldsymbol{y}}_{1:H}) = P(s_1) \prod_{i=1}^{H} \pi(x_i \mid s_i) \mathcal{T}(s_i, x_i, s_{i+1}) \mathbb{1}\{\mathcal{R}(s_i, x_i) = y_i\} \tag{123}$$

The key assumption of a standard MDP is that the transition probability and reward functions depend on the past only through the current state of the system and the action selected by the decision maker in that state. This assumption is called the *Markov property* (Puterman, 1994) and can be characterized using the following independence relationships, for every stage $i = 2, 3, \ldots,$

$$\left(\bar{\boldsymbol{S}}_{1:i-1}, \bar{\boldsymbol{X}}_{1:i-1}, \bar{\boldsymbol{Y}}_{1:i-1} \perp\!\!\!\perp \bar{\boldsymbol{S}}_{i+1:\infty}, \bar{\boldsymbol{X}}_{i:\infty}, \bar{\boldsymbol{Y}}_{i:\infty} \mid S_i\right) \tag{124}$$

In other words, the standard MDP could be seen as a compact representation of a family of joint distributions over observed trajectories of states $\bar{\boldsymbol{S}}_{1:H}$, actions $\bar{\boldsymbol{X}}_{1:H}$, and rewards $\bar{\boldsymbol{Y}}_{1:H}$, provided that the Markov property holds. In the language of structural causality, the Markov property could hold in both the observational distribution $P(\bar{\boldsymbol{s}}_{1:H}, \bar{\boldsymbol{x}}_{1:H}, \bar{\boldsymbol{y}}_{1:H})$ and the interventional distribution $P_\pi(\bar{\boldsymbol{s}}_{1:H}, \bar{\boldsymbol{x}}_{1:H}, \bar{\boldsymbol{y}}_{1:H})$.[21] We showed in Examples 4 and 9 the compression of the observational and interventional distributions of an SCM instance to standard MDP models respectively.

To make the argument more precise, consider the SCM $\mathcal{M}^*$ graphically described in the MDP diagram $\mathcal{G}_{\text{MDP}}$ of Fig. 10d. For every state $i = 1, 2, \ldots,$ conditioning on state S_i and action X_i blocks all paths from history S_j, X_j, Y_j for $j < i$ to any future state S_k, action X_k, and reward Y_k for $k > i$ (Def. 7). The observational distribution evaluated in $\mathcal{M}^*$ thus satisfies the Markov property in Eq. 124 and can be represented using a standard MDP.[22]

Example 21 (MDP, Observational) *Consider the MDP environment $\mathcal{M}^*$ described in Eq. 5. Its observational distribution $P(\bar{\boldsymbol{s}}_{1:H}, \bar{\boldsymbol{x}}_{1:H}, \bar{\boldsymbol{y}}_{1:H})$ defines a standard MDP $\langle \mathscr{D}(S), \mathscr{D}(X), \mathcal{T}_{obs}, \mathcal{R}_{obs} \rangle$ where the transition probability and the reward function are observational quantities given by*

$$\mathcal{T}_{obs}(s, x, s') = P\left(S_{i+1} = s' \mid S_i = s, X_i = x\right) \tag{125}$$

$$\mathcal{R}_{obs}(s, x) = \mathbb{E}\left[Y_i \mid S_i = s, X_i = x\right] \tag{126}$$

Detailed parametrizations of system dynamics $\mathcal{T}$ and $\mathcal{R}$ can be compactly represented as a finite-state machine and are shown in Fig. 5b.

Following the Bellman equation in Eq. 105, we solve for the optimal Q-function $Q_*(s, x)$ in the standard MDP $\langle \mathscr{D}(S), \mathscr{D}(X), \mathcal{T}_{obs}, \mathcal{R}_{obs} \rangle$ and provide it in Table 5. Maximizing the action x for every state s in $Q_*(s, x)$ gives an optimal decision rule $\pi^*_{obs} \triangleq X_i \leftarrow \neg S_i$. ■

21. In this case, the decision rule $\pi(x|s)$ is set as the conditional distribution $P(X_i = x \mid S_i = s)$ defined by the behavioral policy f_X.

22. We will consistently assume that structural functions f_{S_i}, f_{Y_i} and distributions $P(\boldsymbol{U}_{S_i}, \boldsymbol{U}_{Y_i})$ remain invariant across decision horizons $i = 1, 2, \ldots$. This is a common assumption for solving infinite-horizon MDP (Puterman, 1994).

Consider a policy space $\Pi = \{\langle X_i, \{S_i\}\rangle\}_{i=1}^{\infty}$. Following a similar argument, we could show that the interventional distribution induced by any policy $\pi \in \Pi$ evaluated in SCM $\mathcal{M}^*$ satisfies the Markov property, leading to an alternative standard MDP representation.

Example 22 (MDP, Interventional) *Consider the MDP environment $\mathcal{M}^*$ described in Eq. 5. For every policy $\pi = (\pi_i(X_i \mid S_i))_{i=1}^{\infty}$, its interventional distribution $P_\pi(\bar{\boldsymbol{s}}_{1:H}, \bar{\boldsymbol{x}}_{1:H}, \bar{\boldsymbol{y}}_{1:H})$ satisfies the Markov property in Eq. 124. It defines a standard MDP $\langle \mathscr{D}(S), \mathscr{D}(X), \mathcal{T}_{exp}, \mathcal{R}_{exp}\rangle$ where the transition probability and the reward function are interventional quantities given by Eqs. 103 and 104. Their parametrizations are described in the finite-state machine of Fig. 6b.*

We also solve for the optimal Q-function $Q_*(s, x)$ in the MDP $\langle \mathscr{D}(S), \mathscr{D}(X), \mathcal{T}_{exp}, \mathcal{R}_{exp}\rangle$ and obtain an optimal decision rule $\pi^*_{exp} \triangleq X_i \leftarrow S_i$. Revisit Example 16 for detailed computations. ■

Some important observations follow from these two examples. First, the Markov property holds in both the observational and interventional distributions evaluated in the SCM $\mathcal{M}^*$ described in Eq. 5, resulting in two standard MDPs. Second, solving these MDPs leads to different policies π^*_{obs} and π^*_{exp}. The previous discussion in Example 16 showed that only π^*_{exp} is the optimal policy in the underlying environment $\mathcal{M}^*$; while π^*_{obs} is sub-optimal. This suggests that the model assumptions of standard MDPs are generally insufficient in determining the optimal policy in the underlying causal model, however many samples are provided.

Consider now a CRL agent that interacts with the environment $\mathcal{M}^*$ and receives observed data $\mathcal{D}$. Without specifying the learning regime $\mathcal{L}$ (see or do) or causal knowledge $\mathcal{A}$, the agent cannot determine whether data $\mathcal{D}$ is drawn from the observational or interventional distribution from the Markov property. If data $\mathcal{D}$ is collected from passive observations, optimizing the learned standard MDP model could lead to a sub-optimal policy, resulting in unsatisfactory performance. One may wonder if it is possible to recover interventional quantities $\mathcal{T}_{\text{exp}}$ and $\mathcal{R}_{\text{exp}}$ form the observational data $\mathcal{D} \sim P(\boldsymbol{V})$ in MDP environments. Unfortunately, our next result suggests otherwise.

Proposition 1 *For any SCM $\mathcal{M}^*$ compatible with the causal diagram $\mathcal{G}_{\text{MDP}}$ of Fig. 10d, there is an SCM $\mathcal{M}^{(1)}$ compatible with $\mathcal{G}_{\text{MDP}}$ such that for every stage $i = 1, 2, \ldots,$*

$$P^{(1)}\left(s_{i+1} \mid s_i, x_i\right) = P^*\left(s_{i+1} \mid s_i, x_i\right), \qquad \mathbb{E}^{(1)}\left[Y_i \mid s_i, x_i\right] = \mathbb{E}^*\left[Y_i \mid s_i, x_i\right] \tag{127}$$

while

$$P^{(1)}_{x_i}\left(s_{i+1} \mid s_i\right) \neq P^*_{x_i}\left(s_{i+1} \mid s_i\right), \qquad \mathbb{E}^{(1)}_{x_i}\left[Y_i \mid s_i\right] \neq \mathbb{E}^*_{x_i}\left[Y_i \mid s_i\right] \tag{128}$$

■

The following example constructs an alternative SCM $\mathcal{M}^{(1)}$ that generates the observational distribution as the underlying environment $\mathcal{M}^*$, but differs significantly in interventional distributions.

Example 23 (MDP, Observational $\not\Rightarrow$ Interventional) *Consider the following MDP environment*

$$\mathcal{M}^{(1)} = \left\langle \boldsymbol{U} = \{U_{i,1}, U_{i,2}, U_{i,3}\}, \boldsymbol{V} = \{X_i, Y_i, S_i\}, \mathscr{F} = \left\{\mathscr{F}^{(1)}_i\right\}, P^{(1)}(\boldsymbol{U})\right\rangle_{i=1,2,\ldots}, \tag{129}$$

where the causal mechanisms $\mathscr{F}^{(1)}_t$ are defined as

$$\mathscr{F}^{(1)}_t = \begin{cases} S_i \leftarrow S_{i-1} \oplus U_{i-1,2}, \\ X_i \leftarrow S_i \oplus U_{i,1}, \\ Y_i \leftarrow U_{i,3}, \end{cases} \tag{130}$$

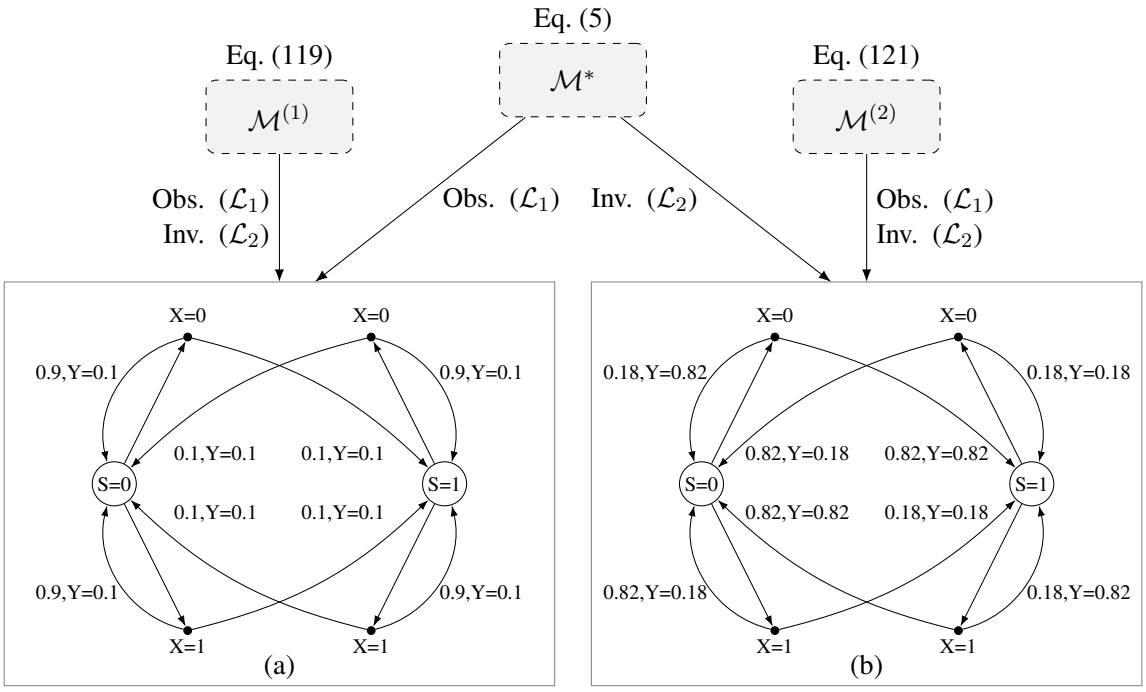


Figure 12: Causal Hierarchy Theorem (CHT) in MDP environments.

and $P^{(1)}(U_{i,1}, U_{i,2}, U_{i,3})$ is such that $U_{i,1}, U_{i,2}, U_{i,3}$ are independent variables drawn from distribution $P(U_{i,1} = 1) = 0.9$, and $P(U_{i,2} = 1) = P(U_{i,3} = 1) = 0.1$.

We compute the observational distributions $P(S_{i+1} \mid S_i, X_i)$ and $\mathbb{E}[Y_i \mid S_i, X_i]$ evaluated in $\mathcal{M}^{(1)}$ and show their parametrization in the finite-state machine in Fig. 12(a). It is verifiable that MDP models $\mathcal{M}^{(1)}$ and $\mathcal{M}^*$ (defined in Eq. 5) generate the same observational distributions ($\mathcal{L}_1$), i.e., the equalities in Eq. 510 hold. On the other hand, we also derive the interventional distribution $P_{X_i}(S_{i+1} \mid S_i)$, $\mathbb{E}_{X_i}[Y \mid S_i]$ evaluated in $\mathcal{M}^{(1)}$. Their parametrization could also be summarized using Fig. 12(a). It follows from previous discussions (Examples 4 and 9) that $\mathcal{M}^{(1)}$ and $\mathcal{M}^*$ (Eq. 5) differ in the interventional distribution ($\mathcal{L}_2$), i.e., inequalities in Eq. 511 hold. ■

The above example shows that the optimal policy in an unknown MDP environment is generally underdetermined by the observational distribution and the Markov property. Conversely, we also show that one cannot recover observational quantities from the interventional data in MDP environments.

Proposition 2 *For any SCM $\mathcal{M}^*$ compatible with the causal diagram $\mathcal{G}_{\text{MDP}}$ of Fig. 10d, there is an SCM $\mathcal{M}^{(2)}$ compatible with $\mathcal{G}_{\text{MDP}}$ such that for every stage $i = 1, 2, \ldots,$*

$$P^{(2)}_{x_i}(s_{i+1} \mid s_i) = P^*_{x_i}(s_{i+1} \mid s_i), \qquad \mathbb{E}^{(2)}_{x_i}[Y_i \mid s_i] = \mathbb{E}^*_{x_i}[Y_i \mid s_i] \tag{131}$$

while

$$P^{(2)}(s_{i+1} \mid s_i, x_i) \neq P^*(s_{i+1} \mid s_i, x_i), \qquad \mathbb{E}^{(2)}[Y_i \mid s_i, x_i] \neq \mathbb{E}^*[Y_i \mid s_i, x_i] \tag{132}$$

■

Our next example corroborates the aforementioned proposition by constructing an alternative SCM $\mathcal{M}^{(2)}$ that generates the same interventional distribution as the underlying environment $\mathcal{M}^*$ but appears different from passive observations.

Example 24 (MDP, Interventional $\not\Rightarrow$ Observational) *Consider the following MDP environment*

$$\mathcal{M}^{(2)} = \left\langle \boldsymbol{U} = \{U_{i,1}, U_{i,2}, U_{i,3}\}, \boldsymbol{V} = \{X_i, Y_i, S_i\}, \mathscr{F} = \left\{\mathscr{F}_i^{(2)}\right\}, P^{(2)}(\boldsymbol{U})\right\rangle_{i=1,2,\dots}, \quad (133)$$

where the causal mechanisms $\mathscr{F}_t^{(2)}$ are defined as:

$$\mathscr{F}_t^{(2)} = \begin{cases} S_i \leftarrow (S_{i-1} \vee X_{t-1}) \oplus U_{i-1,2}, \\ X_i \leftarrow S_i \oplus U_{i,1}, \\ Y_i \leftarrow S_i \oplus X_i \oplus U_{i,3}, \end{cases} \quad (134)$$

and $P^{(2)}(U_{i,1}, U_{i,2}, U_{i,3})$ is such that $U_{i,1}, U_{i,2}, U_{i,3}$ are independent variables drawn from distribution $P(U_{i,1} = 1) = 0.9$, and $P(U_{i,2} = 1) = 0.82 = P(U_{i,3} = 1) = 0.82$.

We compute the interventional distribution $P_{X_i}(S_{i+1} \mid S_i)$ and $\mathbb{E}_{X_i}[Y \mid S_i]$ evaluated in $\mathcal{M}^{(2)}$ and summarize them in the finite-state machine described in Fig. 12(b). It is verifiable that MDP models $\mathcal{M}^{(2)}$ and $\mathcal{M}^*$ (Eq. 5) define the same interventional distributions ($\mathcal{L}_2$), i.e., the equalities in Eq. 526 hold. We also compute the observational distributions $P(S_{i+1} \mid S_i, X_i)$ and $\mathbb{E}[Y \mid S_i, X_i]$ of $\mathcal{M}^{(2)}$ and provide their parametrizations in Fig. 12(b). The previous discussions (Examples 4 and 9) implied that $\mathcal{M}^{(2)}$ and $\mathcal{M}^*$ (Eq. 5) differ significantly in the observational distribution ($\mathcal{L}_1$), i.e., inequalities in Eq. 527 hold. This complements previous examples and illustrates that interventional queries are generally under-determined by observational data in MDP environments. ■

The last two examples are summarized and the results are illustrated in Fig. 12. In the middle of the figure, we show the true MDP model $\mathcal{M}^*$ initially discussed in Example 2 and its induced observational and interventional distributions (described in more detail in Figs. 5 and 6 respectively). Assuming only observational data ($\mathcal{L}_1$) is available, one can construct an alternative SCM $\mathcal{M}^{(1)}$ (left side) that matches the observational distribution but have different interventional behavior (i.e., $\mathcal{L}_1^* = \mathcal{L}_1^{(1)}$, $\mathcal{L}_2^* \neq \mathcal{L}_2^{(1)}$). Formally, the interventional distribution is underdetermined by the observational distribution. Practically, this means that passively observing another agent acting in the environment and collecting samples from it may not be enough to make claims about the agent's policies and their corresponding performance.

On the other hand, the same is the case in the reverse direction; say, whenever interventional data ($\mathcal{L}_2$) is available, one can then construct an alternative SCM $\mathcal{M}^{(2)}$ (right side) that matches the interventional distribution but has a different observational one (i.e., $\mathcal{L}_1^* \neq \mathcal{L}_1^{(2)}$, $\mathcal{L}_2^* = \mathcal{L}_2^{(2)}$). Formally, the observational distribution is underdetermined by the interventional distribution. This may be counter-intuitive since interventions are usually believed to be more informative than just passively observing the system unfold in time. Still, in practice, it doesn't allow the CRL agent to predict how other agents will behave when interacting in the environment. This impossibility will translate into challenges when considering the communication and exchange of experience across agents with the intent of accelerating learning.

More generally, the Causal Hierarchy Theorem (Bareinboim et al., 2020, Thm. 1) states that this impossibility result is strict for almost all causal models. This means it is generically impossible to

draw higher-layer inferences using only lower-layer information. Given that the actual underlying SCM is rarely observable in practice, and no inferences across the layers of the PCH are possible, the CRL agent will need to resort to some causal knowledge and assumptions to make claims about these underlying mechanisms, as discussed in Sec. 2.3. Since the more typical language to describe standard MDPs is constrained to one particular distribution, it's somewhat limiting to consider it as a baseline to the model of the environment/agent relationship given the more general types of tasks that can be represented in terms of the PCH, as discussed earlier in this manuscript.

4. Reinforcement Learning through Causal Lenses

The formalization of causal reinforcement learning tasks (Def. 11) allows us to describe some of the most common and popular learning settings studied in the classic literature of reinforcement learning (RL) and causal inference (CI). This section will investigate learning methods for these classic RL-CI tasks, including off-policy learning (Sec. 4.1), online learning (Sec. 4.2), and causal identification from observational data (Sec. 4.3). These tasks are briefly described in Table 4 and can be seen as variations of the first two dimensions described in the table, i.e., interaction regime and structural assumptions, while the other dimensions are fixed. Lastly, Sec. 4.4 varies these dimensions and moves toward a catalog of novel CRL tasks.

Even though the tasks of off-policy learning, online learning, and causal identification all come from the classic literature, there are still subtle interplays between reinforcement learning and causal invariances that were only formally understood with the CRL formalization. For instance, in the language of structural causality, we will provide a formal justification for off-policy learning algorithms, e.g., inverse propensity weighting and dynamic programming. This permits one to determine when and how to apply RL algorithms to more generalized settings where unobserved confounders exist in the observational dataset. Analyzing these classic tasks through CRL lenses will shed light on other foundational issues and how CI and RL connect.

We will introduce some additional notations before studying these learning tasks in detail. Specifically, we will optimize over a policy space with a finite decision horizon $H = |\boldsymbol{X}| < \infty$. Let actions $\boldsymbol{X}$ be ordered by $X_1, X_2, \ldots, X_H$, $H = |\boldsymbol{X}|$, following a topological order in the underlying SCM $\mathcal{M}^*$. For any indices $i < j$, let $\bar{\boldsymbol{X}}_{i:j} = \{X_i, X_{i+1}, \ldots, X_j\}$ denote a sequence of actions ranging from stage i to stage j. Similarly, let $\bar{\boldsymbol{S}}_{i:j} = \{\boldsymbol{S}_i, \boldsymbol{S}_{i+1}, \ldots, \boldsymbol{S}_j\}$ denote the sequence of states from stage i to stage j. For convenience, we will consistently write $\bar{\boldsymbol{X}}_i = \bar{\boldsymbol{X}}_{1:i}$ and $\bar{\boldsymbol{S}}_i = \bar{\boldsymbol{S}}_{1:i}$. Fix a policy $\pi \in \Pi$. For any indices $i \leq j$, let $(\pi_i, \ldots, \pi_j)$ denote a subsequence of decision rules constrained in π determining values of actions $X_i, \ldots, X_j$.

4.1 Off-Policy Learning

This section investigates the off-policy learning problem where an agent attempts to learn an optimal policy from observational data generated by a different behavior policy (Sutton and Barto, 1998), provided that there is no unmeasured confounder (NUC) in the data (to be defined). First, we will introduce the NUC assumption and discuss how it justifies the off-policy learning approach. We then describe in Sec. 4.1.1 two primary methods in evaluating candidate policies in the off-policy setting, including inverse propensity weighting and dynamic programming (Bellman, 1957).

An off-policy learning agent interacts with the underlying environment (SCM) through passively observing events unfolding over time. Fig. 13 is a graphical representation of the CRL agent interacting with the environment for repeated episodes $t = 1, \ldots, T$. For every episode t, the agent

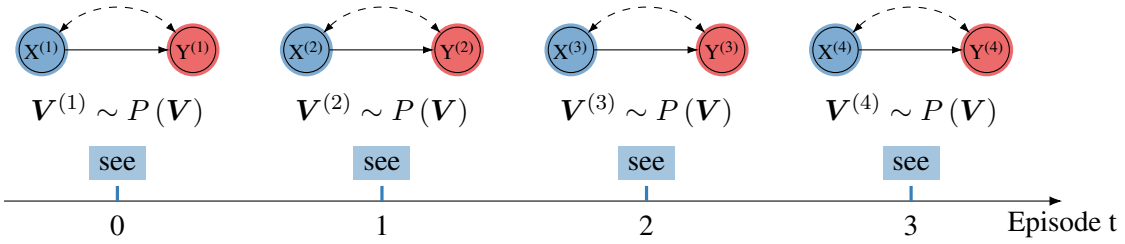


Figure 13: Temporal diagram showing an off-policy learning agent interacting with the environment for repeated episodes.

"sees" the SCM $\mathcal{M}^*$ (Def. 2) and receives an observation $\boldsymbol{V}^{(t)} \sim P(\boldsymbol{V})$. The CRL agent currently aims to use these observations to learn a policy from candidate space Π that maximizes a reward function $\mathcal{R}(\boldsymbol{Y})$. This will be a useful representation to compare other learning modalities and types of interactions. The following signature characterizes this learning setting:

$$\mathcal{T}_{\text{off}} = \left\langle \mathcal{I} = \text{see}, \mathcal{A} = \text{NUC}, \Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}, \mathcal{R} = \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R} \right\rangle. \tag{135}$$

where the agent uses observational data combined with the critical assumption $\mathcal{A} = \text{NUC}$, which means "*No Unmeasured Confounder*". The agent's goal is to obtain an optimal policy estimate from the combination of the observational data and the NUC assumption, i.e.,

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R}(\boldsymbol{Y}) \,\middle|\, \mathcal{A} = \text{NUC},\ \mathcal{D}_{\text{obs}} \sim P(\boldsymbol{V}) \right]. \tag{136}$$

In practice, the NUC assumption will require that at every stage of intervention on action $X_i \in \boldsymbol{X}$, its observed correlations with the reward Y given past actions and covariates' history, is entirely determined by the causal relationships between X_i and Y. In other words, no other variables generate non-causal variations between the decision X_i and the outcome Y. The following definition formalizes this idea.

Definition 13 (No Unmeasured Confounder) *Let $\mathcal{M}^*$ be a SCM and $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}$ be a policy space (Def. 8). The "no unmeasured confounder" (for short, NUC) condition holds if for every action $X_i \in \boldsymbol{X}$, its endogenous parents $\boldsymbol{PA}_i$ and exogenous parents $\boldsymbol{U}_i$ satisfy the following conditions:*

1. *Endogenous parents $\boldsymbol{PA}_i$ are contained in the history $\bar{\boldsymbol{X}}_{i-1} \cup \bar{\boldsymbol{S}}_i$, i.e., $\boldsymbol{PA}_i \subseteq \bar{\boldsymbol{X}}_{i-1} \cup \bar{\boldsymbol{S}}_i$;*
2. *Exogenous parents $\boldsymbol{U}_i$ are independent from exogenous noises $\boldsymbol{U}_j$ associated with all the other endogenous variables V_j in the system, i.e., $\boldsymbol{U}_i \perp\!\!\!\perp \{\boldsymbol{U}_j \mid \forall V_j \in \boldsymbol{V} \setminus \{X_i\}\}$.*

In the above definition, Condition 1 says that all endogenous parents of every action X_i are observed, contained in the past states and actions $\bar{\boldsymbol{S}}_i, \bar{\boldsymbol{X}}_{i-1}$; Condition 2 says that given the past history $\bar{\boldsymbol{X}}_{i-1} \cup \bar{\boldsymbol{S}}_i$, values of every action X_i are decided by an independent noise $\boldsymbol{U}_i$. Note that in the underlying SCM $\mathcal{M}^*$, observational data are generated by a behavior policy $f_{\boldsymbol{X}}$ which determines values of every action X_i based on the endogenous $\boldsymbol{PA}_i$ and exogenous parents $\boldsymbol{U}_i$ for all time steps $i = 1, \ldots, H$. The NUC condition implies that one could simulate the behavior policy using a sequence of decision rules $\left(\pi_i(X_i \mid \bar{\boldsymbol{X}}_{i-1}, \bar{\boldsymbol{S}}_i)\right)_{i=1}^{H}$ such that for every step $i = 1, \ldots, H$,

$\pi_i(X_i \mid \bar{\boldsymbol{X}}_{i-1}, \bar{\boldsymbol{S}}_i) = P(X_i \mid \boldsymbol{PA}_i)$. Allocating actions following these decision rules leads to the following independence relationships, for any sequence of actions $\bar{\boldsymbol{x}}_H$,[23]

$$\left(X_i \perp\!\!\!\perp S_{i+1_{\bar{\boldsymbol{x}}_i}}, \ldots, S_{H_{\bar{\boldsymbol{x}}_{H-1}}}, Y_{\bar{\boldsymbol{x}}_H}, \mid \bar{\boldsymbol{X}}_{i-1}, \bar{\boldsymbol{S}}_i\right) \quad \forall i = 1, \ldots, H \tag{137}$$

Among quantities in the above equation, the potential response $S_{i_{\bar{\boldsymbol{x}}_{i-1}}}, i = 1, \ldots, H$, is the observed state in submodel $\mathcal{M}^*_{\bar{\boldsymbol{x}}_{i-1}}$ induced by the atomic intervention $\text{do}(\bar{\boldsymbol{X}}_{i-1} \leftarrow \bar{\boldsymbol{x}}_{i-1})$; similarly, $Y_{\bar{\boldsymbol{x}}_H}$ is the future potential reward evaluated in submodel $\mathcal{M}^*_{\bar{\boldsymbol{x}}_H}$. The meaning of the NUC condition is illustrated in the next examples.

Example 25 (DTR models where NUC holds) *Consider a 2-stage DTR model $\langle \mathcal{M}^*, \Pi, Y \rangle$ where SCM $\mathcal{M}^*$ is described in Eq. 78 and the policy space $\Pi = \{\langle X_1, \{S_1\}\rangle, \langle X_2, \{S_1, X_1, S_2\}\rangle\}$. We will next examine conditions of Def. 13 and show that they hold in this CDM.*

First, note that for every treatment X_i, $i = 1, 2$, its endogenous parent $\boldsymbol{PA}_i = \{S_i\}$, which is contained in the corresponding input state $\boldsymbol{S}_i$. This implies Condition (1) of NUC holds.

Second, each structural function f_{X_i}, $i = 1, 2$ affecting treatment X_i, the coefficients $\alpha_i = 0$ of the exogenous variable U is equal to zero. This means that there is no unobserved confounder affecting X_i and other variables in the system, i.e., Condition (2) of NUC also holds. Therefore, we conclude the NUC condition holds in the DTR model $\langle \mathcal{M}^*, \Pi, Y \rangle$. ■

Example 26 (DTR models where NUC fails) *Continuing with the DTR model $\langle \mathcal{M}^*, \Pi, Y \rangle$ in the previous example, we now consider an alternative policy space $\Pi' = \{\langle X_1, \emptyset\rangle, \langle X_2, \emptyset\rangle\}$. Every policy $\pi \in \Pi'$ decides values of treatment X_1, X_2 independently, regardless of values of other variables in the system. For $i = 1, 2$, the history $\bar{\boldsymbol{X}}_{i-1} \cup \bar{\boldsymbol{S}}_i = \emptyset$ prior to stage i is an empty set and does not contain the endogenous parent S_i of treatment X_i. Therefore, Condition (1) of NUC fails.*

Alternatively, consider an SCM $\mathcal{M}'$ where for the structural function f_{X_i} of treatment X_i, $i = 1, 2$, the coefficient of the exogenous variable U is equal to $\alpha_i = -3$. This means that there exists an unobserved confounder affecting treatments X_1, X_2 and the primary outcome Y simultaneously. Consequently, Condition (2) of NUC does not hold. ■

4.1.1 OFF-POLICY EVALUATION

Whenever the NUC assumption holds, there exist different strategies that allow one to estimate and compare the effects of candidate policies from observational data without having to perform online experiments in the environment. The first algorithm we discuss that implements this idea is based on a technique known as "*inverse probability weighting*" (IPW) and is widely applied in practice (Rubin, 1974; Robins et al., 2000; Murphy et al., 2001b; Wang et al., 2012; Swaminathan and Joachims, 2015; Liu et al., 2018). Formally,

Theorem 2 (Inverse Propensity Weighting, under NUC) *Let $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ be a CDM where the policy space $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^H$ and the reward function $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$. If NUC holds, for any $\pi \in \Pi$, the expected reward is computable from the observational distribution $P(\boldsymbol{V})$ as*

$$\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})] = \sum_{\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H} \underbrace{\mathbb{E}[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H] P(\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H)}_{\textit{observational distribution}} \underbrace{\prod_{i=1}^{H} \frac{\pi_i(x_i \mid \boldsymbol{s}_i)}{P(x_i \mid \bar{\boldsymbol{x}}_{i-1}, \bar{\boldsymbol{s}}_i)}}_{\textit{ratio } \pi \textit{ and obs. probabilities}}. \tag{138}$$

23. The NUC assumption could also be characterized with a series of graphical conditions based on structural causality, known as *sequential backdoor condition*. We will further elaborate on the graphical implication of NUC in Sec. 4.3.

S_1	X_1	S_2	X_2	Y	$P(s_1, x_1, s_2, x_2, y)$	S_1	X_1	S_2	X_2	Y	$P(s_1, x_1, s_2, x_2, y)$
0	0	0	0	0	0.0128	1	0	0	0	0	0.0042
0	0	0	0	1	0.0466	1	0	0	0	1	0.0011
0	0	0	1	0	0.0009	1	0	0	1	0	0.0011
0	0	0	1	1	0.0585	1	0	0	1	1	0.0042
0	0	1	0	0	0.0013	1	0	1	0	0	0.0004
0	0	1	0	1	0.0049	1	0	1	0	1	0.0001
0	0	1	1	0	0.0269	1	0	1	1	0	0.0098
0	0	1	1	1	0.0982	1	0	1	1	1	0.0027
0	1	0	0	0	0.0442	1	1	0	0	0	0.1013
0	1	0	0	1	0.0121	1	1	0	0	1	0.0008
0	1	0	1	0	0.0008	1	1	0	1	0	0.0796
0	1	0	1	1	0.0554	1	1	0	1	1	0.0218
0	1	1	0	0	0.0051	1	1	1	0	0	0.0130
0	1	1	0	1	0.0014	1	1	1	0	1	0.0001
0	1	1	1	0	0.0281	1	1	1	1	0	0.2566
0	1	1	1	1	0.1028	1	1	1	1	1	0.0040

Table 6: The observational distribution $P(X_1, X_2, S_1, S_2, Y)$ evaluated in the 2-stage DTR environment described in Example 12.

Among the above quantities, $P(x_i \mid \bar{\boldsymbol{x}}_{i-1}, \bar{\boldsymbol{s}}_i)$ measures the natural propensity of the behavior policy for action X_i, $i = 1, \ldots, H$, which is known as the *propensity score*. The IPW estimation requires what is called the *positivity* assumption, i.e., the propensity scores $P(x_i \mid \bar{\boldsymbol{x}}_{i-1}, \bar{\boldsymbol{s}}_i) > 0$ for every entry $\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i$.[24] The following example illustrates the application of the IPW method.

Example 27 *Consider again the CDM $\langle \mathcal{M}^*, \Pi, Y \rangle$ described in Eq. 78 where coefficients $\alpha_1 = \alpha_2 = 0$. We will apply IPW estimation to evaluate the effects of the policy $\pi = (X_1 \leftarrow 0, X_2 \leftarrow 1)$ from the observational distribution $P(S_1, X_1, S_2, X_2, Y)$. Applying the estimation formula provided by Thm. 2 gives*

$$\mathbb{E}^{\text{IPW}}_{X_1 \leftarrow 0, X_2 \leftarrow 1}[Y] = \sum_{s_1, x_1, s_2, x_2} P(s_1, x_1, s_2, x_2, Y = 1) \frac{\mathbb{1}\{x_1 = 0\}}{P(x_1 \mid s_1)} \frac{\mathbb{1}\{x_2 = 1\}}{P(x_2 \mid s_1, x_1, s_2)} \tag{139}$$

24. This quantitative assumption is called *overlap* in (Rosenbaum and Rubin, 1983; Imbens, 2004). There are attempts in the literature to relax it by assuming some parametric models that allow the interpolation of the unobserved areas, e.g., refer to (Rosenbaum, 2002; Kallus and Zhou, 2018).

The detailed parametrization of the observational distribution $P(S_1, X_1, S_2, X_2, Y)$ is provided in Table 6. The above equation could be further written as:

$$\mathbb{E}^{\text{IPW}}_{X_1 \leftarrow 0, X_2 \leftarrow 1}[Y] = \frac{P(S_1 = 0, X_1 = 0, S_2 = 0, X_2 = 1, Y = 1)}{P(X_1 = 0 \mid S_1 = 0)\, P(X_2 = 1 \mid S_1 = 0, X_1 = 0, S_2 = 0)} \tag{140}$$

$$+ \frac{P(S_1 = 0, X_1 = 0, S_2 = 1, X_2 = 1, Y = 1)}{P(X_1 = 0 \mid S_1 = 0)\, P(X_2 = 1 \mid S_1 = 0, X_1 = 0, S_2 = 1)} \tag{141}$$

$$+ \frac{P(S_1 = 1, X_1 = 0, S_2 = 0, X_2 = 1, Y = 1)}{P(X_1 = 0 \mid S_1 = 1)\, P(X_2 = 1 \mid S_1 = 1, X_1 = 0, S_2 = 0)} \tag{142}$$

$$+ \frac{P(S_1 = 1, X_1 = 0, S_2 = 1, X_2 = 1, Y = 1)}{P(X_1 = 0 \mid S_1 = 1)\, P(X_2 = 1 \mid S_1 = 1, X_1 = 0, S_2 = 1)} \tag{143}$$

Evaluating the above equation gives $\mathbb{E}^{\text{IPW}}_{X_1 \leftarrow 0, X_2 \leftarrow 1}[Y] = 0.6757$, which matches the expected reward in Eq. 78, evaluated directly in the SCM $\mathcal{M}^*$. ■

The numeric example above is one instantiation of the larger implication of the theorem showing that the agent does not have to go online and try different actions, but it can learn a policy by simply re-weighting the observational data whenever the conditions of the theorem hold. Interestingly, we further note that by iteratively applying Bayes' rule, the IPW formula in Eq. 138 can be written as

$$\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})] = \sum_{\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H} \mathbb{E}[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H] \prod_{i=1}^{H} P(s_i \mid \bar{\boldsymbol{x}}_{i-1}, \bar{\boldsymbol{s}}_{i-1})\, \pi_i(x_i \mid \boldsymbol{s}_i). \tag{144}$$

Computing the above equation following a reverse topological ordering $i = H, \ldots, 1$ over actions leads to an alternative algorithm for evaluating the effects of candidate policies, based on *dynamic programming* (for short, DP). DP was first introduced in (Bellman, 1957) and has been widely applied in reinforcement learning (Puterman, 1994; Sutton and Barto, 1998). The following proposition describes details for applying DP for off-policy evaluation from the observational distribution, provided that the NUC condition holds.

Theorem 3 (Dynamic Programming) *Let $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ be a CDM where $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}$ and $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$. If NUC holds, for any $\pi \in \Pi$, the expected reward $\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})]$ is computable from the joint distribution $P(\boldsymbol{V})$ as follows:*

$$\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})] = \mathbb{E}\left[\sum_{x_1} Q^{(1)}_\pi(x_1, \boldsymbol{S}_1)\pi_1(x_1 \mid \boldsymbol{S}_1)\right], \tag{145}$$

where the value function $Q^{(i)}_\pi(\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i)$, for $i = 1, \ldots, H-1$, is given by:

$$Q^{(i)}_\pi(\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i) = \mathbb{E}\left[\sum_{x_{i+1}} Q^{(i+1)}_\pi(\bar{\boldsymbol{x}}_{i+1}, \bar{\boldsymbol{s}}_i, \boldsymbol{S}_{i+1})\pi_{i+1}(x_{i+1} \mid \boldsymbol{S}_{i+1}) \,\middle|\, \bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i\right] \tag{146}$$

$$\text{and} \quad Q^{(H)}_\pi(\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H) = \mathbb{E}[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H] \tag{147}$$

■

S_1	X_1	$Q_\pi^{(1)}$	S_1	X_1	$Q_\pi^{(1)}$
0	0	0.8799	1	0	0.4716
0	1	0.8749	1	1	0.1003

(a) $Q_\pi^{(1)}(s_1, x_1)$

S_1	X_1	S_2	X_2	$Q_\pi^{(2)}$	S_1	X_1	S_2	X_2	$Q_*^{(2)}$
0	0	0	0	0.7851	1	0	0	0	0.2149
0	0	0	1	0.9846	1	0	0	1	0.7851
0	0	1	0	0.7851	1	0	1	0	0.2149
0	0	1	1	0.7851	1	0	1	1	0.2149
0	1	0	0	0.2149	1	1	0	0	0.0008
0	1	0	1	0.9846	1	1	0	1	0.2149
0	1	1	0	0.2149	1	1	1	0	0.0008
0	1	1	1	0.7851	1	1	1	1	0.0154

(b) $Q_\pi^{(2)}(s_1, x_1, s_2, x_2)$

Table 7: Evaluation of value functions $Q_\pi^{(1)}(s_1, x_1), Q_\pi^{(2)}(s_1, x_1, s_2, x_2)$ for the policy $\pi = (X_1 \leftarrow 0, X_2 \leftarrow 1)$ in evaluated in 2-stage DTR environment described in Example 12.

It follows from the derivation in Eq. 144 that IPW and DP estimation are, in principle, equivalent. That is, they return the same evaluation for $\mathbb{E}_\pi[Y]$ provided with the same candidate policy π and observational data $P(\boldsymbol{V})$. The following example illustrates this equivalence.

Example 28 *Consider again the CDM $\langle \mathcal{M}^*, \Pi, Y \rangle$ described in Eq. 78 where coefficients $\alpha_1 = \alpha_2 = 0$. We will apply the DP estimation to evaluate the effect of the policy $\pi = (X_1 \leftarrow 0, X_2 \leftarrow 1)$. Thm. 3 allows us to estimate the expected reward $\mathbb{E}_\pi[Y]$ from the observational distribution $P(S_1, X_1, S_2, X_2, Y)$ as follows:*

$$Q_\pi^{(1)}(s_1, x_1) = \sum_{s_2, x_1} Q_\pi^{(2)}(s_1, x_1, s_2, x_2) \mathbb{1}\{x_2 = 1\} P(s_2 \mid x_1, s_1) \tag{148}$$

$$Q_\pi^{(2)}(s_1, x_1, s_2, x_2) = P(Y = 1 | s_1, x_1, s_2, x_2) \tag{149}$$

We compute the parametrization of value functions $Q_\pi^{(1)}(s_1, x_1), Q_\pi^{(2)}(s_1, x_1, s_2, x_2)$ and provide them in Table 7. The expected reward of the policy $\pi = (X_1 \leftarrow 0, X_2 \leftarrow 1)$ is computable as

$$\mathbb{E}^{\text{DP}}_{X_1 \leftarrow 0, X_2 \leftarrow 1}[Y] = \sum_{s_1} Q_\pi^{(1)}(s_1, x_1) \mathbb{1}\{x_1 = 0\} P(s_1) \tag{150}$$

Evaluating the above equation gives $\mathbb{E}^{\text{DP}}_{X_1 \leftarrow 0, X_2 \leftarrow 1}[Y] = 0.6757$, which matches the expected reward in Example 12, evaluated in the SCM $\mathcal{M}^*$. ■

Once IPW and DP evaluation formulas are obtained, efficient methods in the literature estimate the expected rewards of candidate policies from finite samples drawn from the observational distribution $P(\boldsymbol{V})$. For the IPW evaluation, the agent could weigh every observed reward signal with the odds ratio between the target policy π and the propensity score $P(x_i \mid \bar{\boldsymbol{x}}_{i-1}, \bar{\boldsymbol{s}}_i)$, i.e., the second term of Eq. 138. The expected reward is estimable by computing the empirical mean on the weighted rewards. This IPW estimate was first developed to estimate the effects of candidate policies in the single-stage decision setting, i.e., the decision horizon $H = 1$, but later adapted to the problem of estimating the effects of policies in the sequential setting, with the decision horizon $H > 1$. See (Rosenbaum and Rubin, 1983; Robins et al., 2000; Wang et al., 2012; Nahum-Shani et al., 2012) for detailed explanations of how to apply IPW estimation from finite observations provided with the NUC condition.

As for the DP evaluation, the agent first approximates the state-action value function Q in Eq. 145 from the observational data using parametric models and then computes the expected reward of a candidate policy (Tsitsiklis and Van Roy, 1996). For instance, it could approximate Q-functions using a parametric family of linear functions; function parameters are obtainable using the standard least squares regression (Murphy, 2005b). Other more flexible families of parametric models for the Q-functions include regression trees (Ernst et al., 2005), kernels (Ormoneit and Sen, 2002), and neural networks (Mnih et al., 2013). The value function approximation has been studied in the literature under the rubrics of batch reinforcement learning (Bertsekas and Tsitsiklis, 1995; Lange et al., 2012).

When the expected rewards of candidate policies are computable, the agent could then search over the policy space Π and obtain an optimal policy estimate using policy gradient (Sutton et al., 1999). Moreover, when the Q-function $Q_{\pi}^{(i)}(\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i)$ only relies on the state-action value $x_i, \boldsymbol{s}_i$ for every stage of intervention $i = 1, \ldots, H$, one could solve for an optimal policy through iteratively optimizing every decision rule π_i following a reverse topological ordering over actions $\boldsymbol{X}$ (Lauritzen and Nilsson, 2001; Koller and Milch, 2003). This local optimization procedure is analogous to the well-celebrated Q-learning algorithm (Watkins and Dayan, 1992). We refer readers to (Uehara et al., 2022) for a recent literature review on standard off-policy evaluation from finite observational data under the NUC assumption.

Despite the positive results discussed above, IPW and DP methods may fail to recover the effects of candidate policies from observational data whenever the NUC condition (Def. 13) does not hold. The following examples demonstrate this looming challenge.

Example 29 *Consider a MAB model $\langle \mathcal{M}^*, \{\langle X, \emptyset \rangle\}, Y \rangle$ graphically described in Fig. 10a where SCM $\mathcal{M}^*$ is defined in Example 1. In this model, the NUC condition does not hold due to unobserved confounder U affecting action X and reward Y. This implies that IPW is not necessarily applicable to recover the expected reward $\mathbb{E}_x[Y]$ from the observational distribution $P(X, Y)$.*

We will proceed regardless and try to learn a policy $\pi : X \leftarrow 0$. Applying Thm. 2 gives:

$$\mathbb{E}^{\text{IPW}}_{X \leftarrow 0}[Y] = \sum_x \mathbb{E}[Y \mid x] P(x) \frac{\mathbb{1}\{x = 0\}}{P(x)} \tag{151}$$

$$= \frac{P(X = 0, Y = 1)}{P(X = 0)}. \tag{152}$$

Evaluating the above equation gives $\mathbb{E}^{\text{IPW}}_{X \leftarrow 0}[Y] = 0$, which deviates significantly from the actual expected reward $\mathbb{E}_{X \leftarrow 0}[Y] = 0.4$ (Eq. 39) evaluated in SCM $\mathcal{M}^*$.

We also apply DP to evaluate the effect of pulling arm $X \leftarrow 0$, and through Thm. 3, we have:

$$\mathbb{E}^{\text{DP}}_{X\leftarrow 0}[Y] = \sum_x \mathbb{E}[Y \mid x]\, \mathbb{1}\{x = 0\} \tag{153}$$

$$= P(Y = 1 \mid X = 0) \tag{154}$$

Evaluating the above equation gives $\mathbb{E}^{\text{DP}}_{X\leftarrow 0}[Y] = 0$, which, again, deviates from the actual expected reward $\mathbb{E}_{X\leftarrow 0}[Y] = 0.4$ (Eq. 39) evaluated directly in SCM $\mathcal{M}^*$. ■

The above examples show that the validity of off-policy learning methods introduced so far hinges on the NUC assumption. Such a critical assumption could be fragile and does not necessarily hold in many practical settings. For instance, in electronic healthcare records, the physician might prescribe a new drug to patients who are more likely to access high-quality healthcare, thus making the drug appear more effective. For the remainder of this section, we will introduce alternative policy learning assumptions and methods to overcome this issue.

4.2 Online Learning

An online learning agent evaluates candidate policies in space Π by directly deploying them in the underlying environment. A temporal graph illustrating this interaction is described in Fig. 14. In causal language, the agent intervenes in the SCM $\mathcal{M}^*$ for repeated episodes $t = 1, \ldots, T$. For every episode t, it picks a policy $\pi^{(t)} \in \Pi$, performs interventions do $\left(\boldsymbol{X} \leftarrow \pi^{(t)}\right)$ on actions $\boldsymbol{X}$ following $\pi^{(t)}$, and receives subsequent observations $\boldsymbol{V}^{(t)} \sim P_{\pi^{(t)}}(\boldsymbol{V})$.

Formally, an online learning task is described by the following signature:

$$\mathcal{T}_{\text{on}} = \left\langle \mathcal{R} = \text{do}, \mathcal{A} = \emptyset, \Pi = \{\langle X_i, \boldsymbol{S}_i\rangle\}_{i=1}^{H}, \mathcal{R} = \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R} \right\rangle \tag{155}$$

To see the specific optimization in this task, the agent will search for a policy π^* such that

$$\pi^* = \arg\max_{\pi\in\Pi} \mathbb{E}^{\mathcal{M}^*}_{\pi}\left[\mathcal{R}(\boldsymbol{Y}) \;\middle|\; \mathcal{D}_{\text{exp}} \sim P_{\boldsymbol{x}}(\boldsymbol{V})\right], \tag{156}$$

Compared with the off-policy learning task ($\mathcal{T}_{\text{off}}$), an online agent does not make additional structural assumptions about the underlying environment ($\mathcal{A} = \emptyset$), beyond the temporal ordering over state and action variables in the policy space Π (Def. 8). This means that the NUC assumption (Def. 13) discussed earlier does not necessarily hold. Note that for every policy $\pi \in \Pi$, in the submodel $\mathcal{M}^*_\pi$ induced by intervention do(π), all input covariates $\boldsymbol{S}_i$ affecting every action $X_i \in \boldsymbol{X}$ and other variables in the system are observed and measured. This means that the NUC condition is implied in the post-interventional system. The following proposition formalizes this intuition.

Lemma 1 (Experimental NUC) *Let $\langle \mathcal{M}^*, \Pi, \mathcal{R}\rangle$ be a CDM where $\Pi = \{\langle X_i, \boldsymbol{S}_i\rangle\}_{X_i\in\boldsymbol{X}}$ and $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$. For any policy $\pi \in \Pi$, the NUC condition (Def. 13) holds in $\langle \mathcal{M}^*_\pi, \Pi, \mathcal{R}\rangle$ induced by intervention do(π).* ■

Following the NUC condition, the result above can be seen as a formal justification for using standard off-policy learning methods, including IPW and DP, to evaluate other candidate policies $\pi' \in \Pi$ from data drawn from the interventional distribution $P_\pi(\boldsymbol{V})$ through Thms. 2 and 3.

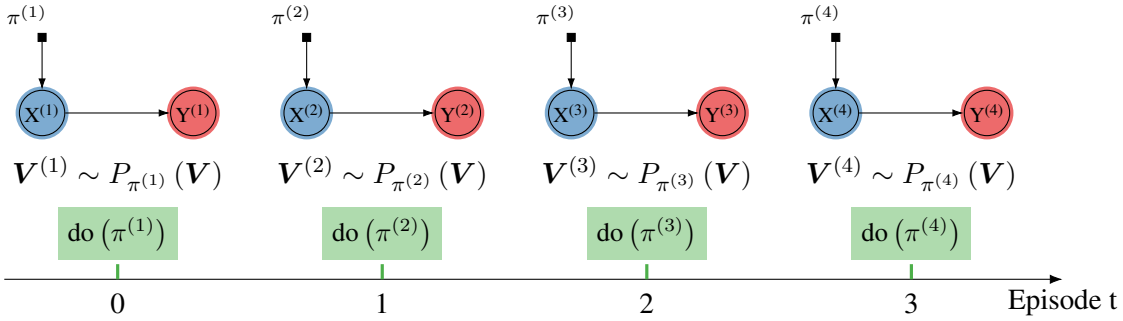


Figure 14: Temporal diagram showing an online learning agent interacting with the environment for repeated episodes.

Example 30 *Consider the 2-stage DTR $\langle \mathcal{M}^*, \Pi, Y \rangle$ where SCM $\mathcal{M}^*$ is described in Eq. 78 with coefficients $\beta_i = -3$; and the policy space $\Pi = \{\langle X_1, \{S_1\}\rangle, \langle X_2, \{S_1, X_1, S_2\}\rangle\}$. It has been shown in Example 26 that the NUC condition does not hold in this model. Now consider an online agent that is deployed in the environment and follows a policy $\pi = (\pi_1, \pi_2)$, where*

$$\pi_i \triangleq \mathbb{1}\{3S_i + U_i > 0\} \tag{157}$$

and U_i, $i = 1, 2$, are independent variable drawn from distribution $\texttt{Logistic}(0, 1)$. Performing interventions do(π) following this policy leads to a submodel described by the following tuple

$$\mathcal{M}^*_\pi = \langle \boldsymbol{U} = \{U, U_1, \ldots, U_5\}, \boldsymbol{V} = \{S_1, X_1, S_2, X_2, Y\}, \mathscr{F}_\pi, P(\boldsymbol{U})\rangle, \tag{158}$$

where the structural functions $\mathscr{F}_\pi$ are given by

$$\mathscr{F}_\pi = \begin{cases} S_1 \leftarrow \mathbb{1}\{U_3 > 0\}, \\ X_1 \leftarrow \mathbb{1}\{3S_1 + U_1 > 0\}, \\ S_2 \leftarrow \mathbb{1}\{0.1 + 0.1S_1 + 0.1X_1 + U_4 > 0\}, \\ X_2 \leftarrow \mathbb{1}\{3S_2 + U_2 > 0\}, \\ Y \leftarrow \mathbb{1}\{3U - 3S_1 - 3X_1 - 3S_1X_1 + 3X_2 - 3S_2X_2 + 3X_1X_2 > 0\}. \end{cases} \tag{159}$$

In the above equations, the unobserved confounder U no longer affects treatments X_1, X_2, and the NUC condition holds in the submodel $\langle \mathcal{M}^*_\pi, \Pi, Y\rangle$. See Example 25 for a detailed discussion. ■

4.2.1 RANDOMIZED CONTROLLED TRIALS

We will discuss different algorithms that systematize the discussion above and operate over the environment in an online fashion. The first algorithm we consider will be called *randomized controlled trials* (for short, RCT) and follows the idea of randomization, which dates back at least to (Fisher, 1935). Fisher's very motivation for considering randomizing the treatment assignment was to eliminate the influence of unmeasured confounders in the collected data.[25] It is a "explore-then-commit" strategy where the agent first explores the environment by determining values of actions

25. Fisher's motivation at the time was to understand the effect of some pesticides on the yield of certain crops (Fisher, 1926). Farmers were biased in how they used pesticides which tended to be applied in the best parts of the land. At

Algorithm 2 Randomized Controlled Trails (RCT)

Require: the policy space Π, the total number of trials $N \in \mathbb{N}$.

for all episodes $t = 1, 2, \ldots$ **do**
  Choose a policy $\pi^{(t)}$ as follows.
  if $t \leq N$ **then**
    Let $\pi^{(t)}$ be a uniform policy

$$\pi_{\text{UNIF}} = (X_1 \sim \texttt{Unif}(\mathscr{D}(X_1)), \ldots, X_H \sim \texttt{Unif}(\mathscr{D}(X_H))) \,. \tag{160}$$

  else
    Let $\pi^{(t)} = \arg\max_{\pi \in \Pi} \hat{\mathbb{E}}^{(N)}_{\pi}[Y]$.
  end if
  Perform $\text{do}(\pi^{(t)})$ for episode t and receive observations $\boldsymbol{V}^{(t)}$.
end for

$\boldsymbol{X}$ uniformly at random for a fixed number of times, and then exploits by committing to a policy that appeared best during exploration.

Alg. 2 shows the detailed experimental design of RCT. It interacts with the underlying environment by repeated episodes of interventions $t = 1, 2, \ldots$. More specifically, for every episode t, RCT selects a policy $\pi^{(t)} \in \Pi$, performs an intervention $\text{do}\left(\pi^{(t)}\right)$, and receives a subsequent observation $V^{(t)} \sim P_{\pi^{(t)}}(\boldsymbol{V})$. During the initialization, the algorithm considers a natural number $N \in \mathbb{N}$, called the *total number of trials*. It determines the total episodes of interventions for the algorithm to explore the environment before committing to a specific policy. For episode $t \leq N$, the algorithm selects a uniform policy π_{UNIF} which determines values of every action $X_i^{(t)}$, $i = 1, \ldots, H$, uniformly at random during the exploration phase. Note that the NUC condition holds under $\text{do}(\pi_{\text{UNIF}})$ intervention (Lem. 1). This means that RCT could compute reward estimates $\hat{\mathbb{E}}^{(N)}_{\pi}[Y]$ for candidate policies $\pi \in \Pi$ from the experimental data $P_{\pi_{\text{UNIF}}}(\boldsymbol{V})$ collected during the exploration. The evaluation procedures were previously described in Thms. 2 and 3. Finally, RCT selects a policy with the highest empirical reward estimates and commits to it for all future episodes $t > N$.[26]

Example 31 *We illustrate RCT in an MAB model $\langle \mathcal{M}^*, \{\langle X, \emptyset \rangle\}, Y \rangle$ where $\mathcal{M}^*$ is an SCM described in Eq. 3 consisting of an arm choice X and reward signal Y. For any policy $\pi(x)$, the expected reward $\mathbb{E}_{\pi}[Y]$ is given by:*

$$\mathbb{E}_{\pi}[Y] = \pi(X = 0)\mathbb{E}_{X \leftarrow 0}[Y] + \pi(X = 1)\mathbb{E}_{X \leftarrow 1}[Y] \tag{161}$$

the end of the season, the pesticides had a higher effect. Fisher was suspicious of this procedure since, in modern terminology, the NUC assumption did not hold. He then had the idea of allocating the treatment randomly to the different plots of land. This insight departed from a tradition led by Pearson (Pearson, 1911) and started a new and fundamental discipline of *experimental design* (Fisher, 1935).

26. The RCT algorithm described in Alg. 2 is also referred to as sequential multiple assignment randomized trials (for short, SMART (Murphy, 2005a)) where every subject (at episode t) is randomized multiple times, one for each stage of decision $X_1, \ldots, X_H$. When the decision horizon $H = 1$, Alg. 2 reduces to the original randomized controlled trials introduced by (Fisher, 1935).

This means that for any stochastic policy $\pi(x) > 0$, $\forall x \in \mathscr{D}(X)$, its performance could always be improved by a deterministic policy $X \leftarrow x^*$ where the optimal arm choice given by[27]

$$x^* = \arg\max_{x \in \{0,1\}} \mathbb{E}_x[Y] \tag{162}$$

It is thus sufficient to estimate the expected reward $\mathbb{E}_x[Y]$ induced by atomic intervention $do(x)$.

Fix the total number of trials N (say, $N = 1,000$). For every episode $t \leq N$, RCT selects an action $X^{(t)}$ uniformly at random over the binary domain $\{0,1\}$, performs an intervention $do\left(X \leftarrow X^{(t)}\right)$, and receives a reward $Y^{(t)} \sim P_{X^{(t)}}(P)$. When the exploration phase is done ($t > N$), RCT estimates the expected reward $\mathbb{E}_x[Y]$ for every action $x \in \{0,1\}$ from finite samples $\left\{X^{(t)}, Y^{(t)}\right\}_{t=1,\dots,N}$. Applying the DP estimation formula (Thm. 3) implies

$$\mathbb{E}_x[Y] = \mathbb{E}_{\pi_{\text{UNIF}}}[Y|x] \tag{163}$$

The empirical reward estimate for an arm x is thus given by

$$\hat{\mathbb{E}}_x^{(N)}[Y] = \frac{1}{N(x)} \sum_{t=1}^{N} Y^{(t)} \mathbb{1}\{X^{(t)} = x\}, \tag{164}$$

where $N(x) = \sum_{t=1}^{N} \mathbb{1}\left\{X^{(t)} = x\right\}$ is the total occurrence of event $X^{(t)} = x$ up to episode N. We could also apply the IPW estimation (Thm. 2) and obtain:

$$\mathbb{E}_x[Y] = \sum_{x'} \mathbb{E}_{\pi_{\text{UNIF}}}\left[Y \mid x'\right] \frac{\mathbb{1}\{x' = x\}}{\pi_{\text{UNIF}}(x)} \tag{165}$$

$$= \mathbb{E}_{\pi_{unif}}\left[Y \frac{\mathbb{1}\{X = x\}}{\pi_{\text{UNIF}}(X)}\right] \tag{166}$$

The last step follows from the definition of expected values. Given samples $\left\{X^{(i)}, Y^{(i)}\right\}_{i=1,\dots,N}$ collected by RCT during exploration ($t < N$), the IPW empirical estimate for the expected reward of pulling arm x is thus given by

$$\hat{\mathbb{E}}_x^{(N)}[Y] = \frac{1}{N} \sum_{t=1}^{N} Y^{(t)} \frac{\mathbb{1}\{X^{(t)} = x\}}{\pi_{\text{UNIF}}\left(X^{(t)}\right)} = \frac{2}{N} \sum_{t=1}^{N} Y^{(t)} \mathbb{1}\{X^{(t)} = x\}. \tag{167}$$

The last step holds since the uniform policy $\pi_{\text{UNIF}}(x) = 1/2$ for any action $x \in \{0,1\}$. The empirical estimates of DP (defined in Eq. 164) and IPW (Eq. 167) coincide if every arm $x \in \{0,1\}$ is equally explored for episodes $t \leq N$, i.e., the total occurrences $N(x) = N/2$. ■

Cumulative Regret We will analyze the performance of RCT algorithm to understand its properties and theoretical guarantees better. Our analysis will focus on an MAB model $\langle \mathcal{M}^*_{\text{MAB}}, \Pi, Y \rangle$, where $\mathcal{M}^*_{\text{MAB}}$ is an MAB environment graphically described in Fig. 10a and the policy space $\Pi = \{\langle X, \emptyset \rangle\}$. Recall that the optimal arm $x^* = \arg\max_x \mathbb{E}_x[Y]$. We denote by $\Delta_x = \mathbb{E}_{x^*}[Y] - \mathbb{E}_x[Y]$

27. It can be shown that in any fixed CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$, the performance of a stochastic policy π could always be improved by a deterministic one. This means that one could optimize the expected reward $\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})]$ over only deterministic policies $\pi \in \Pi$ without loss of generality (Liu and Ihler, 2012, Lem. 2.1).

the gap between the expected reward of playing a suboptimal action $x \in \mathscr{D}(X)$ and the optimal arm x^*. For the analysis' convenience, we also assume that every arm $x \in \mathscr{D}(X)$ is played the same amount of times during the exploration phase, i.e., $N(x) = N/K$ where $K = |\mathscr{D}(X)|$.[28]

There are several ways to measure the performance of online learning algorithms. One popular criterion is to study the algorithm's *cumulative regret* (Auer et al., 2002a), which measures its cumulative loss relative to an optimal strategy that always selects the optimal arm x^*. Formally, the cumulative regret for an online learning algorithm in an MAB environment $\mathcal{M}^*$ after T episodes of trials can be defined as:

$$R(T, \mathcal{M}^*) = \underbrace{T\mathbb{E}_{x^*}\left[Y; \mathcal{M}^*\right]}_{\text{Optimal Reward}} - \underbrace{\sum_{t=1}^{T} \mathbb{E}_{X^{(t)}}\left[Y; \mathcal{M}^*\right]}_{\text{Realized}}. \tag{168}$$

Naturally, minimizing the regret $R(T, \mathcal{M}^*)$ is equivalent to maximizing the total expected reward the agent obtains. A reasonable objective is to design an online learning algorithm that could achieve a sublinear regret, i.e., $R(T, \mathcal{M}^*) = o(T)$.[29] This would imply that its average cumulative regret per episode is converging to zero (Lattimore and Szepesvári, 2020), i.e.,

$$\lim_{T \to \infty} R(T, \mathcal{M}^*)/T = 0 \tag{169}$$

The online learner will eventually close the gap between the optimal strategy that always commits to an optimal arm x^*. In other words, the learner is choosing an optimal arm almost all the time as the total number of episodes T tends to be infinite, i.e.,

$$\lim_{t \to \infty} \mathbb{E}_{X^{(t)}}\left[Y\right] \to \mathbb{E}_{x^*}\left[Y\right] \tag{170}$$

The analysis follows (Lattimore and Szepesvári, 2020, Theorem 6.1). Suppose that there are $\mathscr{D}(X) = \{1, \ldots, K\}$ possible arms. Observe that the `RCT` algorithm only incurs regret in episodes t where it plays a sub-optimal arm x with $\Delta_x > 0$. The cumulative regret after $T > 1$ episodes of interventions could be written as:

$$R(T, \mathcal{M}^*) = \sum_{x: \Delta_x > 0} \Delta_x \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{X^{(t)} = x\right\}\right] \tag{171}$$

In the first N episodes, every arm is played exactly N/K times. Subsequently, it chooses a single action to maximize the empirical reward during exploration. This implies

$$\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{X^{(t)} = x\right\}\right] = \frac{N}{K} + (T - N)P\left(X^{(t)} = x\right) \tag{172}$$

$$\leq \frac{N}{K} + (T - N)P\left(\hat{\mathbb{E}}_x^{(N)}[Y] \neq \max_{x' \neq x} \hat{\mathbb{E}}_{x'}^{(N)}[Y]\right) \tag{173}$$

28. For a uniform policy $\pi_{\texttt{unif}}$, every arm x is expected to be played for $\mathbb{E}[N(x)] = N/K$ on average during exploration.

29. Here we use $\mathcal{O}$ notation, where $f(n) = \mathcal{O}(g(n))$ if function f is bounded above by function g (up to constant factor) asymptotically. That is, $\exists k, \exists n_0$ such that $f(n) \leq kg(n)$ for $\forall n > n_0$. Similarly, $f(n) = o(g(n))$ if $\exists k, \exists n_0$, $f(n) < kg(n)$ for $\forall n > n_0$. For further details on this notation, see (Cormen et al., 2022, Ch. 1.3).

The last step holds since for episodes $t \geq N$, a suboptimal arm $x \neq x^*$ is picked if and only if its empirical reward estimate $\hat{\mathbb{E}}_x^{(N)}[Y]$ is maximal (i.e., the largest). The error probability could thus be bounded by

$$P\left(\hat{\mathbb{E}}_x^{(N)}[Y] \neq \max_{x' \neq x} \hat{\mathbb{E}}_{x'}^{(N)}[Y]\right) \leq P\left(\hat{\mathbb{E}}_x^{(N)}[Y] > \hat{\mathbb{E}}_{x^*}^{(N)}[Y]\right) \tag{174}$$

$$\leq P\left(\hat{\mathbb{E}}_x^{(N)}[Y] - \hat{\mathbb{E}}_{x^*}^{(N)}[Y] - (\mathbb{E}_x[Y] - \mathbb{E}_{x^*}[Y]) > \Delta_x\right) \tag{175}$$

$$\leq \exp\left(-\frac{N\Delta_x^2}{4K}\right) \tag{176}$$

The last step follows the standard concentration inequality (Hoeffding, 1963). Replacing the error probability Eq. 176 into Eq. 173 and summing over regret gives the following bound.

Theorem 4 (Regrets of RCT (Lattimore and Szepesvári, 2020)) *For an MAB $\langle \mathcal{M}^*, \{\langle X, \emptyset\rangle\}, Y\rangle$, let $Y \in \boldsymbol{V}$ be the reward variable with support on $[0,1]$ and let the domain of action X be $\mathscr{D}(X) = \{1, \ldots, K\}$. Fix the total number of trials $N \in \mathbb{N}^+$. The regret of RCT in MAB $\mathcal{M}^*$ after $T > 1$ episodes of interventions is bounded by*

$$R(T, \mathcal{M}^*) \leq \underbrace{\frac{N}{K} \sum_{x:\Delta_x>0} \Delta_x}_{\textit{exploration}} + \underbrace{(T-N) \sum_{x:\Delta_x>0} \Delta_x \exp\left(-\frac{N\Delta_x^2}{4K}\right)}_{\textit{exploitation}} \tag{177}$$

■

The regret bound in Thm. 4 illustrates a trade-off between the exploration and the exploitation stages of the agent's strategy. The first term is the regret cumulated during the exploration ($t \leq N$), and the second term is the expected regret of RCT for picking a suboptimal arm during the exploitation stage. If the total number of trials N is large, the algorithm explores too long, and the regret cumulated during the exploration phase will be large. On the other hand, if N is too small, then the empirical estimate $\hat{\mathbb{E}}_x^{(N)}[Y]$ is more likely to deviate from the expected reward $\mathbb{E}_x[Y]$, and the regret in the exploitation phase increases.

One fundamental question is, therefore, how to choose the optimal number of trials N to balance the amount of exploration versus exploitation. Assume that the number of arms $K = 2$ and the optimal arm $x^* = 1$, and write $\Delta = \Delta_2$. The bound in Eq. 177 simplifies to

$$R(T, \mathcal{M}^*) \leq \frac{N}{2}\Delta + T\Delta \exp\left(-\frac{N\Delta^2}{8}\right). \tag{178}$$

For a large T, the right-hand side of Eq. 178 is minimized up to a rounding error by

$$N = \left\lceil \frac{8}{\Delta^2} \log\left(\frac{T\Delta^2}{4}\right) \right\rceil. \tag{179}$$

For this choice and any $T > 1$, after a few simplifications, the regret of RCT is bounded by

$$R(T, \mathcal{M}^*) \leq \Delta + C\sqrt{T} \tag{180}$$

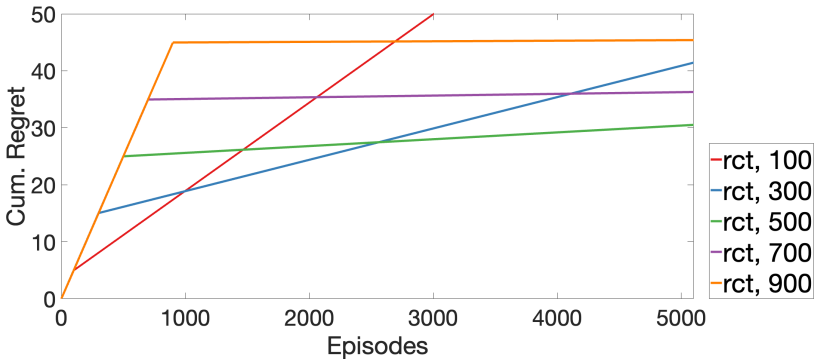


Figure 15: The regret of RCT with varying total number of trials.

where C is a universal constant. That is, RCT is able to achieve a sublinear regret $R(T, \mathcal{M}^*) = \mathcal{O}\left(T^{1/2}\right)$ by fine-tuning the total number of trials N. However, note that the choice of N in Eq. 179 depends on the suboptimal gap Δ and the total number of episodes T, which are not necessarily known in advance. In the next section, we will see an online algorithm that does not depend on the prior knowledge of the model parameter Δ and the total episodes T.

Experiment 1 *Fig. 15 shows the cumulative regrets of RCT when deploying in the MAB model $\langle \mathcal{M}^*, \{\langle X, \emptyset \rangle\}, Y \rangle$ described in Example 31 where the optimal arm choice $X \leftarrow 0$ and the suboptimal gap $\Delta = 0.1$. The optimal arm choice is $X \leftarrow 0$, as shown in the derivation in Example 7.*

Recall that the T represents the total number of episodes that the RCT algorithm interacts with the environment, and the total number trials N represents the amount of exploration it performs (out of T episodes). We evaluate the RCT algorithm with the number of episodes set to $T = 5,000$ and the number of trials set to $N = 100, 300, 500, 700, 900$. Each data point is the average of $1,000$ simulations, which makes the error bars invisible. The simulations show that RCT with $N = 500$ performs the best among all strategies, which is close to the analytical result in Eq. 179, setting $N \approx 878$. This means Eq. 179 provides a near-optimal choice of the total number of trials. ■

4.2.2 THE UPPER CONFIDENCE BOUND ALGORITHM

The upper confidence bound (UCB) algorithm is based on the principle of *optimism in the face of uncertainty* (OFU, Auer et al. 2002b), which states that the agent should act as if the environment is as close to the best-case scenario as possible, given past observations. It offers several advantages over RCT introduced in the previous section, which we summarize below:

- It does not depend on the prior knowledge of the parametrization of the underlying environment, i.e., the gap Δ in the expected rewards between an optimal and a suboptimal policy.
- It does not rely on prior knowledge of the total number of episodes T that the online algorithm will intervene in the underlying environment.
- It achieves the same theoretical guarantees as RCT that fine-tunes the total number of trials N based on prior knowledge of the suboptimal gap Δ and the total episodes T.

The insight of the UCB algorithm is to evoke an *adaptive randomization* strategy (Robbins, 1952; Lai and Robbins, 1985b; Berry and Fristedt, 1985), which means that the agent repeatedly ad-

justs the probability of action assignment according to the past assigned actions and observed outcomes during the experimentation. To make the argument more precise, for an MAB model $\langle \mathcal{M}^*, \{\langle X, \emptyset \rangle\}, Y \rangle$, UCB will utilize the prior actions and rewards' history to compute an **upper confidence bound** to each arm x, which is an overestimate of the unknown expected reward $\mathbb{E}_x[Y]$ with high probability.

In order to better understand the construction of the confidence bound, we need to introduce some basic concentration results. Let $\left\{Y^{(1)}, \ldots, Y^{(n)}\right\}$ be finite i.i.d. reward signals drawn from a discrete distribution $P(Y)$. Let empirical estimates be $\hat{\mathbb{E}}[Y] = \frac{1}{n}\sum_{t=1}^{n} Y^{(i)}$. Applying Hoeffding's inequalities (Hoeffding, 1963) on the reward signal Y bounded in a real interval $[0, 1]$ gives

$$P\left(\mathbb{E}[Y] \geq \hat{\mathbb{E}}[Y] + \sqrt{\frac{\log(1/\delta)}{2n}}\right) \leq \delta \quad \text{for all } \delta \in (0, 1) \tag{181}$$

Fix a sequence of arm selections $x^{(1)}, \ldots, x^{(t)} \in \mathscr{D}(X)$. Let $N_t(x) = \sum_{i=1}^{t} \mathbb{1}\{X^{(t)} = x\}$ be the total occurrence of arm x being played for every $x \in \mathscr{D}(X)$. Given finite samples $\left\{Y^{(1)}, \ldots, Y^{(t)}\right\}$ drawn from interventional distributions $P_{x^{(1)}}(Y), \ldots, P_{x^{(t)}}(Y)$ respectively, it follows from Eq. 181 that the upper confidence bound for the reward $\mathbb{E}_x[Y]$ is defined as[30]

$$\text{UCB}_t(x, \delta) = \underbrace{\hat{\mathbb{E}}_x^{(t)}[Y]}_{exploration} + \underbrace{\sqrt{\frac{\log(1/\delta)}{2N_t(x)}}}_{exploitation} \tag{182}$$

Among quantities in the above equation, the second term is the confidence width computed from the concentration bound in Eq. 181. The first term is the empirical mean estimate $\hat{\mathbb{E}}_x^{(t)}[Y]$ for the expected reward and is given by

$$\hat{\mathbb{E}}_x^{(t)}[Y] = \frac{1}{N_t(x)} \sum_{i=1}^{t} Y^{(i)} \mathbb{1}\left\{X^{(i)} = x\right\} \tag{183}$$

We summarize in Alg. 3 the details of the UCB algorithm when deployed in an unknown MAB model $\mathcal{M}^*_{\text{MAB}}$. For every episode t, it computes confidence bounds $\text{UCB}_{t-1}(x, \delta)$ for every arm $x \in \mathscr{D}(X)$ from prior interventional data $\left\{Y^{(1)}, \ldots, Y^{(t-1)}\right\}$. It then selects an arm $x^{(t)}$ with the maximal upper confidence bound, performs intervention do $\left(x^{(t)}\right)$, and receives subsequent reward $Y^{(t)}$. It can be shown that such an arm allocation strategy based on the upper confidence bound in Eq. 182 balances the trade-off between exploration and exploitation. The algorithm is more likely to play an arm x if it is (1) close to optimal since the empirical reward estimate $\hat{\mathbb{E}}_x^{(t)}[Y]$ is large, or (2) not sufficiently explored and $N_x(t)$ is small. At Step 3, the error probability $\delta = t^{-4}$ decreases as the episode number t increases. This means that the upper confidence bound estimates for every arm x become increasingly accurate as the online learning process continues.

Theorem 5 (Regrets of UCB in MABs (Auer et al., 2002b)) *For an MAB $\langle \mathcal{M}^*, \{\langle X, \emptyset \rangle\}, Y \rangle$, let Y be the reward variable with support on $[0, 1]$, and let the domain of action X be $\mathscr{D}(X) =$*

30. For consistency, we also define $\text{UCB}_t(x, \delta) = \infty$ if $N_t(x) = 0$.

Algorithm 3 Upper Confidence Bound (UCB) in MAB

Require: a policy scope $\Pi = \{\langle X, \emptyset \rangle\}$.
1: **for all** episodes $t = 1, 2, \ldots$ **do**
2: Choose an arm $x^{(t)} = \arg\max_x \text{UCB}_{t-1}(x, \delta)$ where $\delta = t^{-4}$.
3: Perform do $\left(x^{(t)}\right)$ for episode t and receive reward $Y^{(t)}$.
4: **end for**

$\{1, \ldots, K\}$. *It holds the regret of* UCB *in SCM* $\mathcal{M}^*$ *after* $T > 1$ *episodes is bounded by*

$$R(T, \mathcal{M}^*) \leq 8 \sum_{x:\Delta_x>0} \frac{\log(T)}{\Delta_x} + \left(1 + \frac{\pi^2}{3}\right) \sum_{x:\Delta_x>0} \Delta_x \tag{184}$$

■

After a few simplifications (Lattimore and Szepesvári, 2020), the regret bound in Eq. 184 could be further written as:

$$R(T, \mathcal{M}^*) \leq 5 \sum_{x:\Delta_x>0} \Delta_x + C\sqrt{KT\log(T)} \tag{185}$$

where C is a universal constant. In words, UCB achieves a sublinear regret $\mathcal{O}\left(T^{1/2}\log(T)^{1/2}\right)$, which is close to the regret bound obtained by RCT up to logarithmic terms. By employing sharper concentration inequalities for the expected reward estimates, it is possible to shave the dependence on the logarithmic term in the regret bound of UCB (Audibert and Bubeck, 2009). Broadly speaking, the theory supports the claim that UCB is able to overcome the limitation of RCT by removing the dependence on the prior knowledge of suboptimal gaps Δ and the total number of episodes T while achieving the same asymptotic guarantees. Still, in practice, two algorithms' having similar regret bounds does not mean they will perform the same when deployed in the environment. The reason is that the analysis might be loose for one algorithm and not to the other, or by a different margin. For this reason, we now compare the empirical performance of UCB and RCT.

Experiment 2 *Fig. 16 shows the cumulative regret of* UCB *when deploying in the MAB model described in Example 31. The setup is the same as in Experiment 1, which has* $T = 10000$ *and parameter* $\alpha = 0.1$. *As a baseline, we also include* RCT *with various choices for the total number of trials set to* N. *The simulation shows a common phenomenon – if* RCT *is fine-tuned with the optimal choice of the trial number, it can outperform* UCB *by a small margin in the cumulative regret. However, if the trial number* N *must be chosen without prior knowledge of the model parameter* Δ *and the total episodes of interventions* T, UCB *will generally dominate* RCT *in performance.* ■

The principle underpinning UCB has been applied to other RL environments and leads to sublinear regrets, including contextual bandit (Li et al., 2010), Markov decision processes (Auer et al., 2009), and factored MDPs (Osband and Van Roy, 2014), just to name a few. In Sec. 5, we will also see a more generalized implementation of UCB that could find an optimal policy in an arbitrary CRL system.

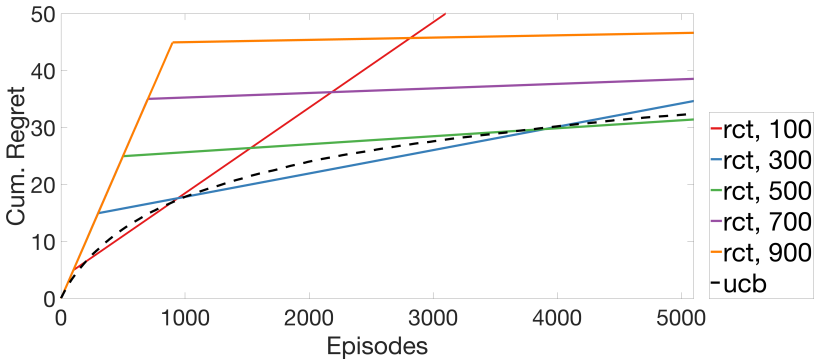


Figure 16: Simulation results comparing performance of online learning algorithms UCB and RCT.

4.3 Causal Identification

Online learning algorithms ensure the NUC condition (Def. 13) holds by deploying candidate policies in the environment (Lem. 1). However, randomized experiments are not applicable in all settings; for example, the effect of a risk factor such as smoking cannot ethically be addressed with randomized controlled trials (Cornfield et al., 1959). Furthermore, performing randomized experiments is expensive, and may even be infeasible due to financial constraints. For instance, in a survey of phase trials of new drugs approved by the Food and Drug Administration (FDA) of the United States from 2015-2016 (Moore et al., 2018), the trial cost was estimated at a median of $41,117 per patient. The increasing cost of certain trials calls for more generalized policy learning methods without direct interventions.[31]

An alternative approach for policy evaluation is to relax the NUC assumption and explore other more general structural assumptions in the underlying environment, represented as a causal diagram. This leads to the setting of causal identification, characterized by a signature:

$$\mathcal{T}_{\text{id}} = \left\langle \mathcal{R} = \text{see}, \mathcal{A} = \mathcal{G}, \Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}, \mathcal{R} = \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R} \right\rangle \tag{186}$$

The optimization in this task goes as follows:

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R}\left(\boldsymbol{Y}\right) \ \middle| \mathcal{A} = \mathcal{G},\ \mathcal{D}_{\text{obs}} \sim P(\boldsymbol{V}) \right], \tag{187}$$

Similar to off-policy learning previously discussed in Sec. 4.1, a causal identification agent collected data from the underlying SCM $\mathcal{M}^*$ through repeated episodes of passive observations ($\mathcal{I} = \text{see}$). or every episode t, the agent observes the SCM $\mathcal{M}^*$, and receives a sample drawn from the observational distribution $P(\boldsymbol{V})$. The key difference is that the agent does not assume the NUC condition (Def. 13). Instead, it now has access to a causal diagram $\mathcal{G}$ associated with the underlying environment $\mathcal{M}^*$ (Def. 6).[32][33] It will then incorporate experts' domain knowledge to obtain a causal

31. The situation is more challenging in practice since even in settings where RCTs are applicable, there are still serious concerns due to issues of transportability (also known in the literature as external validity/generalizability). For further discussion, refer to (Bareinboim and Pearl, 2016; Correa and Bareinboim, 2020b).
32. There exist causal discovery algorithms to learn (an equivalence class of) causal diagrams from observational (Spirtes et al., 2000; Pearl, 2000) and experimental data (Kocaoglu et al., 2017, 2019; Jaber et al., 2020).
33. This settings can be generalized to consider input distributions in which the other agent is known to have collected data under a randomized regime, albeit of another action variable different than $\boldsymbol{X}$, say $\boldsymbol{Z}$. This problem has been studied in the literature under the rubric of g-identification (Bareinboim and Pearl, 2012a; Lee et al., 2019).

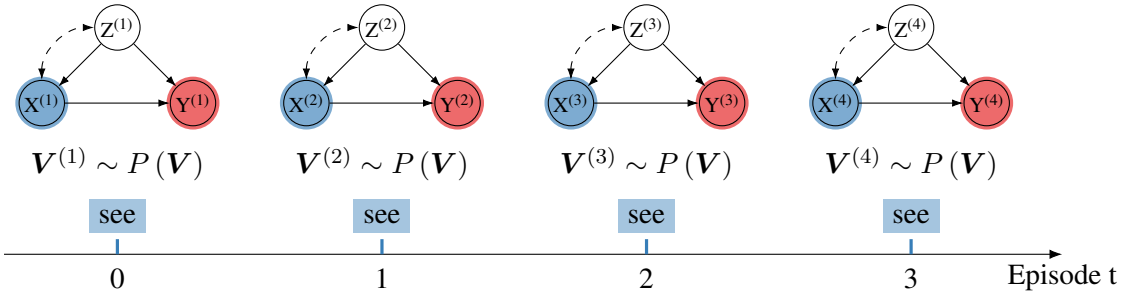


Figure 17: Temporal diagram showing a causal identification agent interacting with the environment for repeated episodes.

diagram compatible with the underlying environment. See Sec. 2.3 for more discussion. This interaction is illustrated in the temporal diagram in Fig. 17.

The agent aims to learn an optimal policy $\pi^* \in \Pi$ from the combination of the observational data $P(\boldsymbol{V})$ and structural assumptions encoded in the causal diagram $\mathcal{G}$. A key challenge is to find a function of the observational distribution $P(\boldsymbol{V})$ that is guaranteed to be equal to the probability query of interest in the intervened submodel $\mathcal{M}_\pi$, for any SCM $\mathcal{M}$ compatible with structural assumptions encoded in the diagram $\mathcal{G}$. We introduce next this notion articulated more formally.

Definition 14 (Identifiability) *Let subsets of variables $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{V}$ and π be a policy over $\boldsymbol{X}$. The interventional distribution $P_\pi(\boldsymbol{Y})$ is identifiable from the structural assumptions $\mathcal{A}$ if $P_\pi(\boldsymbol{Y})$ is uniquely computable from any positive observational distribution $P(\boldsymbol{V})$ in any SCM $\mathcal{M}$ satisfying $\mathcal{A}$. That is, if for every pair of SCMs $\mathcal{M}_1$ and $\mathcal{M}_2$ compatible with structural assumptions $\mathcal{A}$, $P_\pi(\boldsymbol{Y}; \mathcal{M}_1) = P_\pi(\boldsymbol{Y}; \mathcal{M}_2)$ whenever $P(\boldsymbol{V}; \mathcal{M}_1) = P(\boldsymbol{V}; \mathcal{M}_2) > 0$.*

For a causal identification task, the structural assumptions $\mathcal{A}$ will be encoded as a causal diagram $\mathcal{G}$. We say an interventional distribution $P_\pi(\boldsymbol{Y})$ is identifiable from a causal diagram $\mathcal{G}$ if for any pair of SCMs $\mathcal{M}_1, \mathcal{M}_2$, $P_\pi(\boldsymbol{Y}; \mathcal{M}_1) = P_\pi(\boldsymbol{Y}; \mathcal{M}_2)$ whenever $P(\boldsymbol{V}; \mathcal{M}_1) = P(\boldsymbol{V}; \mathcal{M}_2) > 0$ and $\mathcal{G}(\mathcal{M}_1) = \mathcal{G}(\mathcal{M}_2) = \mathcal{G}$. On the other hand, as for an off-policy learning task described in Sec. 4.1, the structural assumptions $\mathcal{A}$ are specified using the NUC condition (Def. 13), which restricts the form of the behavior policy $f_{\boldsymbol{X}}$ determining values of actions $\boldsymbol{X}$ in the underlying environment. It follows as a corollary from Thms. 2 and 3 that for any policy $\pi \in \Pi$, the expected reward $\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})]$ is identifiable provided with the NUC condition (Def. 13).

Corollary 1 *For a policy space $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^H$ and reward function $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$, let a policy $\pi \in \Pi$. $P_\pi(\boldsymbol{Y})$ is identifiable from the NUC condition (Def. 13) w.r.t. the policy space Π.*

The following example shows the identifiability guarantee from the NUC assumption.

Example 32 (Identification under NUC) *We consider a 2-stage DTR $\langle \mathcal{M}^*, \Pi, Y \rangle$ described in Fig. 10c where the policy space $\Pi = \{\langle X_1, \{S_1\} \rangle, \langle X_2, \{S_1, X_1, S_2\} \rangle\}$. We further assume that the NUC condition (Def. 13) holds, i.e., the unobserved confounder U does not affect actions X_1, X_2. This means that for every action X_1 (or X_2), when its direct parents S_1 (or S_1, X_1, S_2) are observed and measured, its values are only affected by independent noise.*

It follows from IPW estimation of Thm. 2 that the expected reward of any policy $\pi \in \Pi$ is computable from the observational distribution and policy's definition, and is given by:

$$\mathbb{E}_\pi[Y] = \sum_{s_1,x_1,s_2,s_2} \underbrace{\frac{\mathbb{E}[Y \mid s_1,x_1,s_2,s_2]\, P(s_1,x_1,s_2,s_2)}{P(x_1|s_1)P(x_2|s_1,x_1,s_2)}}_{\textit{observational distribution}} \underbrace{\pi_1(x_1|s_1)\pi_1(x_2|s_1,x_1,s_2)}_{\textit{policy } \pi}. \tag{188}$$

The above equation is a function of the candidate policy π (the second term) and the observational distribution $P(S_1, X_1, S_2, X_2, Y)$ (the first term). Formally, the expected reward $\mathbb{E}_\pi[Y]$ is identifiable from $P(\boldsymbol{V})$ provided with the NUC condition, i.e., no matter the specific form of the mechanisms $\boldsymbol{F}$ and exogenous conditions $P(\boldsymbol{u})$ of the underlying, true SCM $\mathcal{M}^*$. We also compute the value function of policy $\pi \in \Pi$ following the DP estimation in Thm. 3:

$$Q_\pi^{(1)}(x_1,s_1) = \sum_{s_2,x_2} Q_\pi^{(2)}(x_1,x_2,s_1,s_2)P(s_2|s_1,x_1)\pi_2(x_2 \mid s_1,x_1,s_2) \tag{189}$$

$$Q_\pi^{(2)}(x_1,x_2,s_1,s_2) = \mathbb{E}[Y|s_1,x_1,s_2,s_2]. \tag{190}$$

Finally, the expected reward is identifiable by

$$\mathbb{E}_\pi[Y] = \sum_{s_1,x_1} Q_\pi^{(1)}(x_1,s_1)\pi_1(x_1 \mid s_1)P(s_1), \tag{191}$$

and value functions $Q_\pi^{(1)}, Q_\pi^{(2)}$ are both computable from distribution $P(S_1, X_1, S_2, X_2, Y)$. ■

4.3.1 SEQUENTIAL BACKDOOR FOR POLICY EVALUATION

As previously discussed in Sec. 4.1, both IPW (Thm. 2) and DP (Thm. 3) are popular off-policy evaluation algorithms in causal inference and reinforcement learning literature. Therefore, it is worth understanding the conditions under which these algorithms are applicable for identifying the expected rewards of policies in a policy space Π, provided with an arbitrary causal diagram $\mathcal{G}$. There exists a graphical condition called he *sequential backdoor* (Pearl and Robins, 1995) that delimits whether the effect of performing a sequence of atomic interventions can be identified by covariates adjustments. In this section, we generalize the sequential backdoor criterion to evaluate the effects of sequential policies (i.e., not necessarily atomic[34]), which select actions based on values of other observed covariates in the system.

Before describing the details of the criterion, we first introduce some necessary notations. For every policy $\pi \in \Pi$, let $\mathcal{G}_\pi$ denote the causal diagram associated with the submodel $\mathcal{M}_\pi^*$ induced by policy intervention do(π). Operationally, the manipulated diagram $\mathcal{G}_\pi$ is obtained from $\mathcal{G}$ by performing the following procedures: for every $i = 1, \ldots, H$,

1. Remove all incoming arrows pointing into action node X_i;

2. Add arrows from nodes in input states $\boldsymbol{S}_i$ to the action node X_i.

34. For a more nuanced and detailed discussion on the difference between atomic and non-atomic interventions, please refer to (Correa and Bareinboim, 2020a, Appendix B).

For a policy $\pi = (\pi_1, \ldots, \pi_H)$, let $(\pi_i, \ldots, \pi_j)$ be a sub-policy consisting of decision rules with restriction to indices $1 \leq i < j \leq H$. The manipulated graph $\mathcal{G}_{\pi_i,\ldots,\pi_j}$ is thus a causal diagram obtained by intervention do $(\pi_i, \ldots, \pi_j)$ following the sub-policy $(\pi_i, \ldots, \pi_j)$.[35] As an example, consider the causal diagram $\mathcal{G}$ described Fig. 18a. For a policy $(\pi_1(X_1 \mid S_1), \pi_2(X_2 \mid S_1, X_1, S_2))$, Fig. 18b shows the manipulated causal diagram $\mathcal{G}_{\pi_1,\pi_2}$ induced by intervention do(π_1, π_2); the added incoming arrows to actions X_1, X_2 are highlighted in blue. The diagram $\mathcal{G}_{\pi_2}$ induced by intervention do(π_2) following a sub-policy with restriction on action X_2 is shown in Fig. 18c. Formally, the sequential backdoor condition for identifying policy interventions is defined as follows.

Definition 15 (Sequential Backdoor Condition (Policy Intervention)) *Let $\mathcal{G}$ be causal diagram and $\boldsymbol{Y} \subseteq \boldsymbol{V}$ be reward signals. A policy space $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}$ is said to satisfy the* sequential backdoor condition *w.r.t. $\boldsymbol{Y}$ in $\mathcal{G}$ (for short, Π is backdoor admissible) if for every policy $\pi \in \Pi$, the following condition hold: for every $i = 1, \ldots, H$,*

$$\left(X_i \perp\!\!\!\perp \boldsymbol{Y} \mid \bar{\boldsymbol{X}}_{i-1}, \bar{\boldsymbol{S}}_i\right)_{\mathcal{G}_{\underline{X_i},\pi_{i+1},\ldots,\pi_H}} \tag{192}$$

That is, conditioning on nodes $\bar{\boldsymbol{X}}_{i-1} \cup \bar{\boldsymbol{S}}_i$ d-separates all backdoor paths *from action X_i to reward signals $\boldsymbol{Y}$ that contains an arrow pointing into X_i in the manipulated graph $\mathcal{G}_{\pi_{i+1},\ldots,\pi_H}$.*

In spirit, the independence relationship in Eq. 192 is similar to the celebrated backdoor criterion (Pearl, 2000, Def. 3.3.1). A backdoor path between nodes $\boldsymbol{X}$ and $\boldsymbol{Y}$ is a sequence of edges starting with an arrow pointing into a node in $\boldsymbol{X}$. This criterion ensures that at every stage $i = 1, \ldots, H$, the actions and covariates' history $\bar{\boldsymbol{X}}_{i-1}, \bar{\boldsymbol{S}}_i$ effectively summarize all the information that the behavior policy uses to determine values of action X_i, which are also relevant to the reward signal Y. In other words, all the confounders between action X_i and reward Y are measured, and no other variables could generate non-causal correlations between X_i and Y.

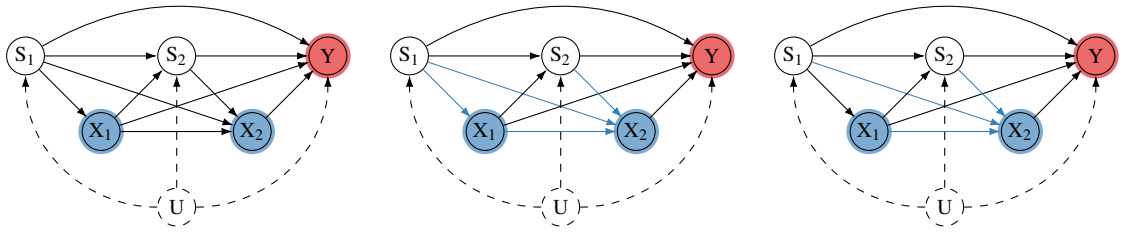
As a special case, let $\pi = (\pi_1, \ldots, \pi_H)$ be an atomic policy such that every decision rule $\pi_i \triangleq X_i \leftarrow x_i$ for every step $i = 1, \ldots, H$. In this case, the manipulated graph $\mathcal{G}_{\pi_{i+1},\ldots,\pi_H}$, $i = 1, \ldots, H$, is a subgraph obtained from $\mathcal{G}$ by removing the incoming arrows of every action node $X_{i+1}, \ldots, X_H$. The independence relationship in Eq. 192 reduces to the sequential backdoor condition for atomic interventions do$(\boldsymbol{x})$ in (Pearl and Robins, 1995) and could be written as:

$$\left(X_i \perp\!\!\!\perp \boldsymbol{Y} \mid \bar{\boldsymbol{X}}_{i-1}, \bar{\boldsymbol{S}}_i\right)_{\mathcal{G}_{\underline{X_i},\overline{X_{i+1},\ldots,X_H}}}. \tag{193}$$

Note that in the above equation, $\mathcal{G}_{\underline{X_i},\overline{X_{i+1},\ldots,X_H}}$ is a subgraph contained in the manipulated diagram $\mathcal{G}_{\underline{X_i},\pi_{i+1},\ldots,\pi_H}$ (w.r.t. an arbitrary policy π over actions $\boldsymbol{X}$). This means that the independence condition in Eq. 192 is stronger than the one in Eq. 193. The atomic backdoor condition in (Pearl and Robins, 1995) holds whenever the policy space Π is backdoor admissible in diagram $\mathcal{G}$.

Whenever the NUC condition (Def. 13) holds, note that the variables $\bar{\boldsymbol{X}}_{i-1}, \bar{\boldsymbol{S}}_i$ contain all direct parents of action X_i. There is no unobserved confounder affecting X_i and any other variable in the environment. Conditioning on the actions and states' history $\bar{\boldsymbol{X}}_{i-1}, \bar{\boldsymbol{S}}_i$ thus "blocks" all backdoor path from action node X_i to reward nodes $\boldsymbol{Y}$. As a corollary, it follows immediately that Def. 15 subsumes the NUC condition (Def. 13).

35. Note that for $i = 1, \ldots, H$, covariates $\boldsymbol{S}_i$ are non-descendent of actions $\bar{\boldsymbol{X}}_{i:H}$. It is verifiable that for every policy $\pi \in \Pi$, the manipulated graph $\mathcal{G}_{\pi_{i+1},\ldots,\pi_H}$, $i = 1 \ldots, H$, is acyclic and preserves the topological ordering $\boldsymbol{S}_1 \prec X_1 \prec \cdots \prec \boldsymbol{S}_H \prec X_H$.



(a) $\mathcal{G}$ (b) $\mathcal{G}_{\pi_1,\pi_2}$ (c) $\mathcal{G}_{\pi_2}$

Figure 18: Causal diagram satisfying the NUC condition and its manipulated diagrams.

Corollary 2 *For a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ where $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{X_i \in \boldsymbol{X}}$ and $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$, let $\mathcal{G}$ be the causal diagram associated with SCM $\mathcal{M}^*$. If the NUC holds in $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$, then Π is backdoor admissible w.r.t reward signals $\boldsymbol{Y}$ in diagram $\mathcal{G}$.*

Whenever the NUC condition holds in a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$, the above corollary implies that the policy space Π must satisfy the sequential backdoor with regard to reward signals $\boldsymbol{Y}$ in the causal diagram $\mathcal{G}$ of the underlying SCM $\mathcal{M}^*$. The following example demonstrates this intuition.

Example 33 (NUC $\Rightarrow$ Sequential Backdoor) *Consider a 2-stage DTR model $\langle \mathcal{M}^*, \Pi, Y \rangle$ consisting of actions X_1, X_2, observed states S_1, S_2, and a primary outcome Y; the policy space $\Pi = \{\langle X_1, \{S_1\}\rangle, \langle X_2, \{S_1, X_1, S_2\}\rangle\}$. Assume that the NUC condition holds (Def. 13), which means that the unobserved confounder U does not affect actions X_1, X_2. Fig. 18a shows a more detailed causal diagram $\mathcal{G}$ associated with the environment $\mathcal{M}^*$; the incoming arrows $U \to X_1$, $U \to X_2$ are now removed.*

We will next show that the policy space Π satisfies sequential backdoor condition with regard to the primary outcome Y in the causal diagram $\mathcal{G}$. Consider first the action X_1. For every policy $(\pi_1, \pi_2) \in \Pi$, the manipulated diagram $\mathcal{G}_{\pi_2}$ is shown in Fig. 18c. One could see by inspection that conditioning on covariate S_1 d-separates all backdoor paths from X_1 to Y in $\mathcal{G}_{\pi_2}$, i.e.,

$$(X_1 \perp\!\!\!\perp Y \mid S_1)_{\mathcal{G}_{\underline{X_1},\pi_2}} \tag{194}$$

We also examine the independence relationship of Eq. 192 with regard to X_2, which is the last action in the decision sequence. It is thus sufficient to consider the causal diagram $\mathcal{G}$ of Fig. 18a. Again, conditioning input covariates S_1, X_1, S_2 d-separates backdoor paths from X_2 to Y in $\mathcal{G}$, i.e.,

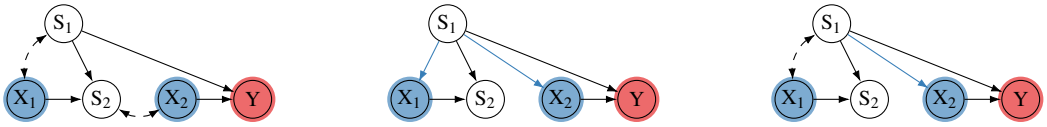
$$(X_2 \perp\!\!\!\perp Y \mid S_1, X_1, S_2)_{\mathcal{G}_{\underline{X_2}}} \tag{195}$$

We thus conclude the policy space Π is backdoor-admissible w.r.t. the primary outcome Y in $\mathcal{G}$. ■

Interestingly, Def. 15 covers more general conditions than the NUC. There are CDMs $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ where the NUC does not hold while the policy space Π satisfies the sequential backdoor condition in the causal diagram $\mathcal{G}$ associated with the SCM $\mathcal{M}^*$.

Example 34 (Sequential Backdoor $\not\Rightarrow$ NUC) *More formally, consider an SCM*

$$\mathcal{M}^* = \left\langle \boldsymbol{U} = \{U_i\}_{i=1}^5, \boldsymbol{V} = \{S_1, S_2, X_1, X_2, Y\}, \mathscr{F}, P(\boldsymbol{U}) \right\rangle \tag{196}$$



(a) $\mathcal{G}$ (b) $\mathcal{G}_{\pi_1,\pi_2}$ (c) $\mathcal{G}_{\pi_2}$

Figure 19: Causal diagram satisfying the sequential backdoor and its manipulated diagrams.

consisting of actions X_1, X_2 and a reward signal Y. The causal mechanisms $\mathscr{F}$ are defined as:

$$\mathscr{F} = \begin{cases} S_1 \leftarrow U_1, \\ X_1 \leftarrow U_1 \oplus U_4, \\ S_2 \leftarrow S_1 \oplus X_1 \oplus U_2, \\ X_2 \leftarrow U_2 \oplus U_5, \\ Y \leftarrow S_1 \oplus X_1 \oplus U_3, \end{cases} \tag{197}$$

The exogenous distribution $P(\boldsymbol{U})$ is defined such that U_1, U_2, U_3 are independent binary variables following distribution $P(U_i = 0) = 0.9$, $i = 1, \ldots, 3$; also, U_4, U_5 are independent noise uniformly drawn over $\{0, 1\}$. Fig. 19a shows the causal diagram $\mathcal{G}$ associated with the SCM $\mathcal{M}^*$.

Consider a policy space $\Pi = \{\langle X_1, \{S_1\}\rangle, \langle X_2, \{S_1\}\rangle\}$. One could see by inspection that the NUC condition does not hold in the CRL system $\langle \mathcal{M}^*, \Pi, Y\rangle$ due to the presence of unobserved confounders U_i, $i = 1, 2$, affecting action (X_i) and state (S_i) variables, simultaneously. We next examine the scope Π and see if it satisfies the sequential backdoor criterion in the diagram $\mathcal{G}$. For a policy $(\pi_1, \pi_2) \in \Pi$, the manipulated diagram $\mathcal{G}_{\pi_2}$ is shown in Fig. 19c. Conditioning on the covariate S_1 d-separates the backdoor path $X_1 \dashleftarrow\dashrightarrow S_1 \rightarrow Y$ in $\mathcal{G}_{\pi_2}$, i.e.,

$$(X_1 \perp\!\!\!\perp Y \mid S_1)_{\mathcal{G}_{\underline{X_1},\pi_2}} \tag{198}$$

Similarly, conditioning on S_1 also blocks the backdoor path $X_2 \dashleftarrow\dashrightarrow S_2 \leftarrow X_1 \dashleftarrow\dashrightarrow S_1 \rightarrow Y$ between action X_2 and reward Y, i.e.,

$$(X_2 \perp\!\!\!\perp Y \mid S_1)_{\mathcal{G}_{\underline{X_2}}} \tag{199}$$

The above independence relationships imply Π satisfies the sequential backdoor w.r.t. the reward Y in the diagram $\mathcal{G}$ even when the NUC does not hold in the decision model $\langle \mathcal{M}^*, \Pi, Y\rangle$. ■

One important observation here is that the NUC condition is hard to ascertain unless implied by the (physical) randomization procedure or assumptions following the model, such as the ones required by the sequential back-door condition. This means that despite the NUC condition being popular throughout the literature, the same is not a primitive but a byproduct of more fundamental notions. Our next result establishes the identifiability of the effects of candidate policies in Π, provided that the policy scope Π is backdoor-admissible in the causal diagram $\mathcal{G}$.

Theorem 6 *Let $\mathcal{G}$ be a causal diagram, $\Pi = \{\langle X_i, \boldsymbol{S}_i\rangle\}_{i=1}^{H}$ be a policy space, and $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$ be a reward function. If Π is backdoor-admissible w.r.t. $\boldsymbol{Y}$ in $\mathcal{G}$ (Def. 15), for any policy $\pi \in \Pi$, the expected reward $\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})]$ is identifiable from $\mathcal{G}$. Moreover, $\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})]$ is computable from the observational distribution $P(\boldsymbol{V})$ following the IPW (Thm. 2) or the DP (Thm. 3) estimation.* ■

S_1	X_1	X_2	Y	$P(s_1, x_1, x_2, y)$	S_1	X_1	X_2	Y	$P(s_1, x_1, x_2, y)$
0	0	0	0	0.2025	1	0	0	0	0.0025
0	0	0	1	0.0225	1	0	0	1	0.0225
0	0	1	0	0.0225	1	0	1	0	0.0225
0	0	1	1	0.2025	1	0	1	1	0.0025
0	1	0	0	0.2025	1	1	0	0	0.0025
0	1	0	1	0.0225	1	1	0	1	0.0225
0	1	1	0	0.0225	1	1	1	0	0.0225
0	1	1	1	0.2025	1	1	1	1	0.0025

Table 8: The observational distribution $P(S_1, X_1, X_2, Y)$ of the SCM $\mathcal{M}^*$ defined in Eq. 197.

In words, the sequential backdoor condition in Def. 15 generalizes standard off-policy learning methods to settings where the NUC does not hold, and there exist unobserved confounders affecting actions and other variables in the system. As long as the sequential backdoor holds, one could apply IPW and DP algorithms to evaluate candidate policies from the observational data while ascertaining the validity of the estimation procedure in the limit.

Example 35 *Consider again the SCM $\mathcal{M}^*$ described in Eq. 197. We are interested in evaluating a policy $\pi = (\pi_1, \pi_2)$ such that $\pi_1 : X_1 \leftarrow S_1, \pi_2 : X_2 \leftarrow \neg S_1$. The submodel entailed by intervention do(π_1, π_2) is described by the tuple*

$$\mathcal{M}^*_\pi = \left\langle \boldsymbol{U} = \{U_i\}_{i=1}^5, \boldsymbol{V} = \{S_1, X_1, S_2, X_2, Y\}, \mathscr{F}_\pi, P(\boldsymbol{U}) \right\rangle, \tag{200}$$

where the structural functions $\mathscr{F}_\pi$ is given by

$$\mathscr{F}_\pi = \begin{cases} S_1 \leftarrow U_1, \\ X_1 \leftarrow S_1, \\ S_2 \leftarrow S_1 \oplus X_1 \oplus U_2, \\ X_2 \leftarrow \neg S_1, \\ Y \leftarrow S_1 \oplus X_1 \oplus U_3, \end{cases} \tag{201}$$

Evaluating the expected reward Y in submodel $\mathcal{M}^*_\pi$ results in

$$\begin{aligned} \mathbb{E}_{X_1 \leftarrow S_1, X_2 \leftarrow \neg S_1}[Y] &= \mathbb{E}[S_1 \oplus \neg S_1 \oplus U_3] && (202) \\ &= P(0 \oplus U_3 = 1) && (203) \end{aligned}$$

Evaluating the above equation gives $\mathbb{E}_{X_1 \leftarrow S_1, X_2 \leftarrow \neg S_1}[Y] = 0.9$. Since the agent does not have access to $\mathcal{M}^*$, we apply next the IPW estimation procedure (Thm. 2) to evaluate the effects of policy $\pi = (X_1 \leftarrow S_1, X_2 \leftarrow \neg S_1)$ from the observational distribution. Recall that scope $\Pi = \{\langle X_1, \{S_1\}\rangle, \langle X_2, \{S_1\}\rangle\}$ is backdoor-admissible in the causal diagram $\mathcal{G}$ of Fig. 18a. Thm. 6

S_1	X_1	$Q_\pi^{(1)}$	S_1	X_1	$Q_\pi^{(1)}$
0	0	0.9	1	0	0.9
0	1	0.9	1	1	0.9

(a) $Q_\pi^{(1)}(s_1, x_1)$

S_1	X_1	X_2	$Q_\pi^{(2)}$	S_1	X_1	X_2	$Q_\pi^{(2)}$
0	0	0	0.1	1	0	0	0.9
0	0	1	0.9	1	0	1	0.1
0	1	0	0.1	1	1	0	0.9
0	1	1	0.9	1	1	1	0.1

(b) $Q_\pi^{(2)}(s_1, x_1, x_2)$

Table 9: Evaluation of value functions $Q_\pi^{(1)}, Q_\pi^{(2)}$ for policy $\pi = (X_1 \leftarrow S_1, X_2 \leftarrow \neg S_1)$ in SCM $\mathcal{M}^*$ described in Eq. 197.

allows us to compute the expected reward from $P(S_1, X_1, X_2, Y)$ as follows:

$$\begin{aligned}\mathbb{E}^{\text{IPW}}_{X_1 \leftarrow S_1, X_2 \leftarrow \neg S_1}[Y] &= \sum_{s_1, x_1, x_2} P(s_1, x_1, x_2, Y = 1) \frac{\mathbb{1}\{x_1 = s_1\}}{P'(x_1 \mid s_1)} \frac{\mathbb{1}\{x_2 = \neg s_1\}}{P(x_2 \mid s_1, x_1)} \\ &= 4 \sum_{s_1, x_1, x_2} P(s_1, x_1, x_2, Y = 1) \mathbb{1}\{x_1 = s_1, x_2 = \neg s_1\} \end{aligned} \tag{204}$$

The last step holds since $X_1 \leftarrow U_1 \oplus U_4$, $X_2 \leftarrow U_2 \oplus U_5$; and U_4, U_5 are independent noise uniformly drawn over $\{0, 1\}$. The complete parametrization for the observational distribution $P(S_1, X_1, X_2, Y)$ is provided in Table 8. The above equation could thus be further written as:

$$\begin{aligned}\mathbb{E}^{\text{IPW}}_{X_1 \leftarrow S_1, X_2 \leftarrow \neg S_2}[Y] &= 4P(S_1 = 0, X_1 = 0, X_2 = 1, Y = 1) \\ &\quad + 4P'(S_1 = 1, X_1 = 1, X_2 = 0, Y = 1) \end{aligned} \tag{205}$$

The above evaluation matches the expected reward in Eq. 203, evaluated in SCM $\mathcal{M}^*$. ■

Example 36 *We also apply the DP estimation (Thm. 3) to evaluate the effect of policy $\pi = (X_1 \leftarrow S_1, X_2 \leftarrow \neg S_1)$ in SCM $\mathcal{M}^*$, described in Eq. 197. By applying Thm. 6, we obtain the expected reward $\mathbb{E}_\pi[Y]$ from the observational distribution $P(S_1, X_1, X_2, Y)$ as follows:*

$$Q_\pi^{(1)}(s_1, x_1) = \sum_{x_2} \mathbb{1}\{x_2 = \neg s_1\} Q_\pi^{(2)}(s_1, x_1, x_2) \tag{206}$$

$$Q_\pi^{(2)}(s_1, x_1, x_2) = P(Y = 1 | s_1, x_1, x_2) \tag{207}$$

We compute the above value functions $Q_\pi^{(1)}(s_1, x_1), Q_\pi^{(2)}(s_1, x_1, x_2)$; their parameterizations are provided in Table 9. Finally, the expected reward is identifiable by

$$\begin{aligned}\mathbb{E}^{\text{DP}}_{X_1 \leftarrow S_1, X_2 \leftarrow \neg S_1}[Y] &= \sum_{s_1, x_1} \mathbb{1}\{x_1 = s_1\} Q_\pi^{(1)}(s_1, x_1) P(s_1) \\ &= Q_\pi^{(1)}(0, 0) P(S_1 = 0) + Q_\pi^{(1)}(1, 1) P(S_1 = 1) \end{aligned} \tag{208}$$

The above computation matches the expected reward in Eq. 203, evaluated in SCM $\mathcal{M}^*$. ■

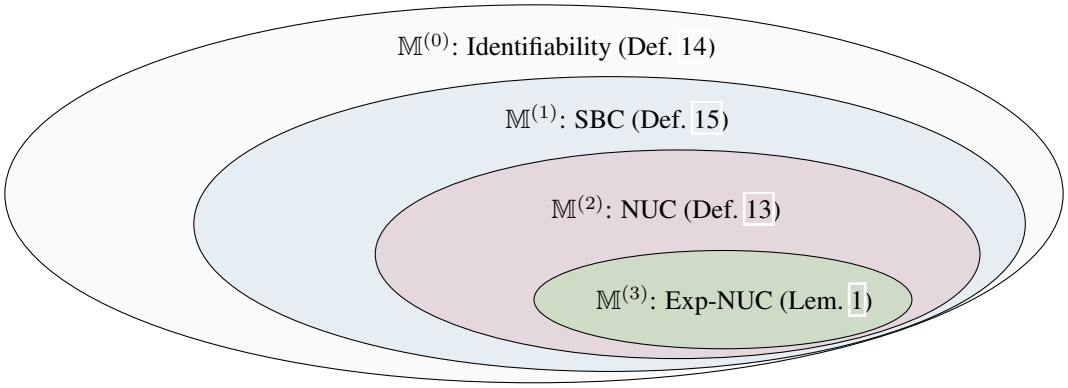


Figure 20: Assumptions under which IPW and DP algorithms are applicable

We provide in Fig. 20 a summary of the relationships of structural assumptions about the data-generating process discussed so far that permit the evaluation of the effects of candidate policies.

- Sec. 4.1 provides a sufficient assumption called the NUC (Def. 13) that allows one to evaluate policy effects from the observational data via the application of IPW and DP algorithms (Thms. 2 and 3). The family of SCMs satisfying the NUC condition is denoted by $\mathbb{M}^{(2)}$, which is marked in red in the figure, and is the baseline of most of the current literature. However, the NUC assumption is opaque and not testable in practice, which may lead to potentially wrong inferences about the optimal policy (as shown in Examples 29).

- In Sec. 4.2, we show that the NUC condition holds in all submodels induced by interventions $\text{do}(\pi)$ following candidate policies $\pi \in \Pi$, which is called Exp-NUC. This family of submodels is denoted by $\mathbb{M}^{(3)}$ and is marked in dark green in the figure. In words, whenever an agent goes online in the environment and collects experimental data following a known policy, the same data can be used to evaluate the effect of new policies, as implied by Lemma 1. Formally, this implies the NUC condition and constitutes a sufficient condition for the identification of the effects of policy interventions.

- On the other side in blue, Sec. 4.3.1 provides a more generalized graphical condition, called the sequential backdoor criterion (SBC, Def. 15). The family of SCMs satisfying the sequential backdoor condition is denoted by $\mathbb{M}^{(1)}$. This graphical condition is defined based on the causal diagram encoding the underlying causal mechanisms, which could be more readily evaluatable from the available data and by domain experts. It is sufficient in determining whether IPW and DP are applicable to evaluate candidate policies from observational data.

- Finally, the outer ellipse $\mathbb{M}^{(0)}$ describes the set of all environments (SCMs) where the effects of candidate policies are identifiable from the underlying data-generating mechanisms.

After all, Fig. 20 shows that $\mathbb{M}^{(3)} \subset \mathbb{M}^{(2)} \subset \mathbb{M}^{(1)}$; and this containment relationship is strict. This means that there exists a causal model that is not induced by the intervention $\text{do}(\pi)$ and satisfies the NUC condition (Example 25). There also exists a causal model where the NUC does not hold, but is backdoor admissible (Example 34). One may wonder if the containment $\mathbb{M}^{(1)} \subset \mathbb{M}^{(0)}$ is also

strict. We will show next that this is the case. Particularly, there are learning settings where the sequential backdoor criterion does not hold, but the agent could still explore the structural constraints in the causal diagram to recover the effects of candidate policies from the observational data.

4.3.2 DO-CALCULUS LEARNING

Our discussion begins with an example illustrating policy evaluation from observational data under a set of non-parametric constraints known as the front-door (Pearl, 2000, Sec. 3.1.2).

Example 37 (Front-door Environment) *Consider the causal diagram $\mathcal{G}$ described in Fig. 21a, which is also known as the "front-door" diagram. We are interested in evaluating the expected reward $\mathbb{E}_x[Y]$ of an atomic policy $\pi : X \leftarrow x$ that sets the value of action X to a constant x. Due to the presence of a backdoor path $X \dashleftarrow\dashrightarrow Y$, the policy space $\Pi = \{\langle X, \emptyset\rangle\}$ does not satisfy the backdoor criterion (Def. 15), and standard off-policy learning algorithm do not generally apply.*

In our context, this means that the policy from which the data is coming from was implemented in the environment where the agent had access to the unobserved confounder, marked as the dashed-bidirected arrow in the graph. This unobserved confounder seems to suggest that the expected reward $\mathbb{E}_x[Y]$ is not identifiable from the Front-Door diagram, and the agent should, therefore, go online following the discussion in Sec. 4.2.

However, existing results in causal inference suggest otherwise. By applying the Front-Door adjustment in (Pearl, 2000, Thm. 3.3.4), the expected reward $\mathbb{E}_x[Y]$ can be computed from the observational distribution $P(X, Y, W)$ through the following mapping:

$$\mathbb{E}_x[Y] = \sum_w P(w \mid x) \sum_{x'} \mathbb{E}\left[Y \mid w, x'\right] P(x'). \tag{209}$$

Among the quantities in the above equation, $P(w \mid x)$, $\mathbb{E}[Y \mid w, x']$, and $P(x')$ are all functions of the observational distribution $P(X, Y, W)$. Therefore, the expected reward $\mathbb{E}_x[Y]$ is identifiable from the front-door diagram. The learner could then evaluate the expected reward of every arm $X \leftarrow x$ from the observational data and solve for the optimal treatment x^*. ■

For the remainder of this section, we will introduce complete machinery to identify the expected rewards of candidate policies $\pi \in \Pi$ from the causal diagram $\mathcal{G}$. Our discussion begins with a procedure to reduce the original identification problem into identifying effects of atomic interventions (e.g., do($\boldsymbol{x}$)) from the same diagram $\mathcal{G}$. Let $\boldsymbol{Y} \subseteq \boldsymbol{V}$ be an arbitrary subset of endogenous variables. For any policy $\pi \in \Pi$, evaluating the joint distribution over $\boldsymbol{Y}$ in submodel $\mathcal{M}_\pi$ is given by:

$$P_\pi(\boldsymbol{y}) = \sum_{\boldsymbol{v}\setminus\boldsymbol{y}} \sum_{\boldsymbol{u}} P(\boldsymbol{u}) \prod_{V \in \boldsymbol{V}\setminus\boldsymbol{X}} P(v \mid \boldsymbol{pa}_V, \boldsymbol{u}_V) \prod_{X_i \in \boldsymbol{X}} \pi_i(x_i \mid \boldsymbol{s}_i) \tag{210}$$

Recall that $\mathcal{G}_\pi$ is a manipulated graph obtained from $\mathcal{G}$ by replacing incoming arrows of node X_i with arrows from covariates $\boldsymbol{S}_i$ to X_i for every action $X_i \in \boldsymbol{X}$. Let $\boldsymbol{Z} = An(\boldsymbol{Y})_{\mathcal{G}_\pi}$ be ancestors of

nodes $\boldsymbol{Y}$. Then all the non-ancestor nodes can be summed out from the above equation leading to[36]

$$P_\pi(\boldsymbol{y}) = \sum_{\boldsymbol{z}\setminus\boldsymbol{y}} \sum_{\boldsymbol{u}} P(\boldsymbol{u}) \prod_{V\in\boldsymbol{Z}\setminus\boldsymbol{X}} P(v \mid \boldsymbol{pa}_V, \boldsymbol{u}_V) \prod_{X_i\in\boldsymbol{X}\cap\boldsymbol{Z}} \pi_i(x_i \mid \boldsymbol{s}_i) \tag{211}$$

$$= \sum_{\boldsymbol{x}^*,\boldsymbol{s}^*} \underbrace{P_{\boldsymbol{x}^*}(\boldsymbol{y},\boldsymbol{s}^*)}_{\text{atomic intervention}} \prod_{X_i\in\boldsymbol{X}^*} \underbrace{\pi_i(x_i \mid \boldsymbol{s}_i)}_{\text{new policy } \pi} \tag{212}$$

where $\boldsymbol{X}^* = \boldsymbol{X} \cap \boldsymbol{Z}$ are actions in $\boldsymbol{X}$ that are ancestors of $\boldsymbol{Y}$ in the post-interventional graph $\mathcal{G}_\pi$; $\boldsymbol{S}^* = \boldsymbol{Z} \setminus (\boldsymbol{X} \cup \boldsymbol{Y})$ are ancestors of $\boldsymbol{Y}$ in $\mathcal{G}_\pi$, excluding $\boldsymbol{Y}$ and $\boldsymbol{X}$. It follows from Eq. 212 that the interventional distribution $P_\pi(\boldsymbol{Y})$ is identifiable if the distribution $P_{\boldsymbol{x}^*}(\boldsymbol{Y}, \boldsymbol{S}^*)$ induced by atomic intervention do$(\boldsymbol{x}^*)$ is identifiable. The following proposition implies that the reverse also holds.

Proposition 3 (Correa and Bareinboim (2019); Tian (2004)) *Let $\mathcal{G}$ be a causal diagram, Π be a policy space $\{\langle X_i, \boldsymbol{S}\rangle\}_{i=1}^H$, and $\boldsymbol{Y} \subseteq \boldsymbol{V}$ be a set of variables. For any policy $\pi \in \Pi$, $P_\pi(\boldsymbol{Y})$ is identifiable from $\mathcal{G}$ if and only if $P_{\boldsymbol{x}^*}(\boldsymbol{Y}, \boldsymbol{S}^*)$ for any $\boldsymbol{x}^* \in \mathscr{D}(\boldsymbol{X}^*)$ is identifiable from $\mathcal{G}$.* ■

The following examples demonstrate the decomposition in Eq. 212 with various causal diagrams.

Example 38 *Consider the Front-door diagram $\mathcal{G}$ described in Fig. 21a and a policy scope $\Pi = \{\langle X, \emptyset\rangle\}$. For any policy $\pi(X)$, the post-interventional graph $\mathcal{G}_\pi$ is a chain $X \to W \to Y$. For the reward Y in this graph, its ancestor action $\boldsymbol{X}^* = An(Y) \cap \{X\} = \{X\}$ and other ancestor nodes $\boldsymbol{S}^* = An(Y) \setminus \{Y, X\} = \{W\}$. Following the decomposition in Eq. 212, the expected reward $\mathbb{E}_\pi[Y]$ for any policy $\pi \in \Pi$ could be written as*

$$\mathbb{E}_\pi[Y] = \sum_y y P_\pi(y)$$

$$= \sum_y y \sum_{x,w} P_x(y, w)\, \pi(x) \tag{213}$$

Among quantities in the above equation, $P_x(Y, W)$ is the interventional distribution induced by atomic intervention do$(X \leftarrow x)$. ■

Example 39 *Figs. 18a and 18b show a causal diagram $\mathcal{G}$ and the post-interventional diagram $\mathcal{G}_\pi$ associated with a policy space $\Pi = \{\langle X_1, \{S_1\}\rangle, \langle X_2, \{S_1\}\rangle\}$. For the reward node Y in graph $\mathcal{G}_\pi$, ancestor actions $\boldsymbol{X}^* = An(Y) \cap \{X_1, X_2\} = \{X_2\}$; and covariates $\boldsymbol{S}^* = An(Y) \setminus \{Y, X_1, X_2\} = \{S_1\}$. For any policy $\pi \in \Pi$, the expected reward $\mathbb{E}_\pi[Y]$ could be written as:*

$$\mathbb{E}_\pi[Y] = \sum_y y P_\pi(y)$$

$$= \sum_y y \sum_{x_2,s_1} P_{x_2}(y, s_1)\, \pi_2(x_2|s_1) \tag{214}$$

The last step follows from the decomposition of Eq. 212. Among quantities in the above equation, $P_{x_2}(Y, S_1)$ is the interventional distribution induced by atomic intervention do$(X_2 \leftarrow x_2)$. ■

36. The decomposition in Eq. 212 was introduced in (Tian, 2004) and extended for identifying policy effects from out-of-domain distributions in (Correa and Bareinboim, 2020b) .

Lem. 3 implies that in order to evaluate candidate policies from the observational distribution, it is sufficient to identify the corresponding effects induced by atomic interventions. Such a problem has been studied in the literature, and several algorithms and graphical criteria have been proposed (Pearl, 2000; Spirtes et al., 2001). First and foremost, we formally introduce do-calculus (Pearl, 1995), which consists of three inferential rules. Each rule dictates that two interventional distributions are equivalent under a condition that can be read off from the causal diagram corresponding to the underlying, unobserved SCM.

Theorem 7 (Rules of do-calculus (Pearl, 2000)) *Let $\mathcal{G}$ be a causal diagram compatible with a structural causal model $\mathcal{M}$, with endogenous variables $\boldsymbol{V}$. For any disjoint subsets $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W} \subseteq \boldsymbol{V}$, the following rules hold for interventional distributions compatible with $\mathcal{G}$:*

Rule 1 *Insertion/deletion of observations:*

$$P_{\boldsymbol{x}}(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{w}) = P_{\boldsymbol{x}}(\boldsymbol{y} \mid \boldsymbol{w}) \text{ if } (\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{W}) \text{ in } \mathcal{G}_{\overline{\boldsymbol{X}}} \tag{215}$$

Rule 2 *Action/observation exchange:*

$$P_{\boldsymbol{x},\boldsymbol{z}}(\boldsymbol{y} \mid \boldsymbol{w}) = P_{\boldsymbol{x}}(\boldsymbol{y} \mid \boldsymbol{z}, \boldsymbol{w}) \text{ if } (\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{W}) \text{ in } \mathcal{G}_{\overline{\boldsymbol{X}}\underline{\boldsymbol{Z}}} \tag{216}$$

Rule 3 *Insertion/deletion of actions:*

$$P_{\boldsymbol{x},\boldsymbol{z}}(\boldsymbol{y} \mid \boldsymbol{w}) = P_{\boldsymbol{x}}(\boldsymbol{y} \mid \boldsymbol{w}) \text{ if } (\boldsymbol{Y} \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X}, \boldsymbol{W}) \text{ in } \mathcal{G}_{\overline{\boldsymbol{X}}\overline{\boldsymbol{Z(W)}}} \tag{217}$$

where $\boldsymbol{Z(W)}$ is the subset of nodes in $\boldsymbol{Z}$ that are not ancestors of $\boldsymbol{W}$-nodes in $\mathcal{G}_{\overline{\boldsymbol{X}}}$ ■

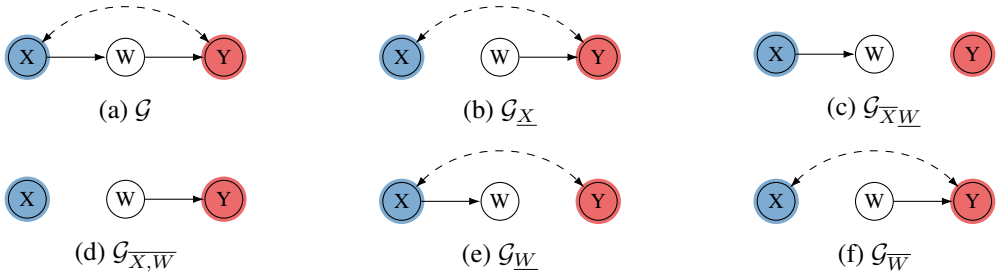
The first rule affirms that the d-separation criterion also holds for causal diagrams under intervention. The second rule gives the condition for when observing and intervening on variables $\boldsymbol{Z}$ are equivalent from the perspective of outcomes $\boldsymbol{Y}$. The third rule gives the conditions for when the do-operator can be removed entirely from the expression, i.e., there is no causal effect of $\boldsymbol{Z}$ on $\boldsymbol{Y}$.

We call *do-calculus learning* an algorithmic procedure to identify causal effects from the observational distribution through the applications of do-calculus together with standard mathematical rules, or in some equivalent, perhaps more systematic form (Tian and Pearl, 2002b). Such a procedure can be shown sufficient and necessary to identify causal effects from observational (Shpitser and Pearl, 2006a; Huang and Valtorta, 2006b) and interventional distributions (Bareinboim and Pearl, 2012b; Lee et al., 2019). This means that if $P_{\boldsymbol{x}}(\boldsymbol{Y})$ cannot be expressed in terms of observational probabilities $P(\boldsymbol{V})$ (or $P_{\boldsymbol{z}}(\boldsymbol{V})$, for some $\boldsymbol{Z}$) by repeated applications of these three rules together with basic probability algebra, such an expression does not exist, and the effect is non-identifiable.

Proposition 4 *Rules of Do-calculus, together with standard probability manipulations, are sound and complete for determining the identifiability of all interventional distributions of the form $P_{\boldsymbol{x}}(\boldsymbol{Y})$ from a causal diagram $\mathcal{G}$ and the available observational and interventional distributions.* ■

The following examples demonstrate how to apply do-calculus learning to evaluate candidate policies with an arbitrary policy space Π from the observational distribution in different causal diagrams.

Example 40 *Consider the front-door diagram $\mathcal{G}$ in Fig. 21a. The decomposition in Eq. 213 implies that in order to evaluate the effect of a policy $\pi(X)$, it is sufficient to identify the interventional distribution $P_x(Y, W)$. We try to remove the subscript from every probability term that appears in the*



(a) $\mathcal{G}$ (b) $\mathcal{G}_{\underline{X}}$ (c) $\mathcal{G}_{\overline{X}\underline{W}}$

(d) $\mathcal{G}_{\overline{X,W}}$ (e) $\mathcal{G}_{\underline{W}}$ (f) $\mathcal{G}_{\overline{W}}$

Figure 21: A front-door graph and its manipulated representations.

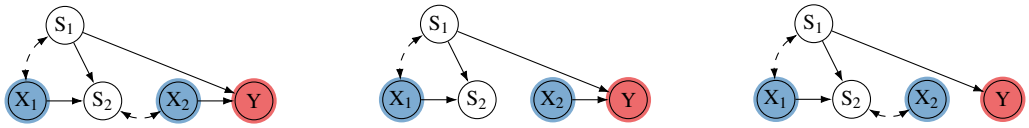
expression of $P_x(Y, W)$ since its absence represents the fact that the causal effect is expressible in terms of the observational distribution, hence computable from the available data (independent from the underlying functions and exogenous variables). A derivation of a 'subscript-free' expression for $P_x(Y, W)$ is given below Eqs. 218 and 224. We illustrate the application of rules of do-calculus, Eqs. 219-224, in Figs. 21b-21f.

$$
\begin{aligned}
P_x(y, w) &= P_x(y|w)P_x(w) && \textit{Probability Axioms} && (218)\\
&= P_x(y|w)P(w|x) && \textit{Rule 2 } (W \perp\!\!\!\perp X)_{\mathcal{G}_{\underline{X}}} && (219)\\
&= P_{x,w}(y)P(w|x) && \textit{Rule 2 } (Y \perp\!\!\!\perp W \mid X)_{\mathcal{G}_{\overline{X}\underline{W}}} && (220)\\
&= P_w(y)P(w|x) && \textit{Rule 3 } (Y \perp\!\!\!\perp X \mid W)_{\mathcal{G}_{\overline{X,W}}} && (221)\\
&= P(w|x)\sum_{x'} P_w(y|x')P_w(x') && \textit{Probability Axioms} && (222)\\
&= P(w|x)\sum_{x'} P(y|w, x')P_w(x') && \textit{Rule 2 } (Y \perp\!\!\!\perp W \mid X)_{\mathcal{G}_{\underline{W}}} && (223)\\
&= P(w|x)\sum_{x'} P(y|w, x')P(x') && \textit{Rule 3 } (X \perp\!\!\!\perp W)_{\mathcal{G}_{\overline{W}}} && (224)
\end{aligned}
$$

We note that the do-operator (i.e., the subscript) does not appear in the final expression in Eq. 224, so even though we do not possess any quantitative knowledge about the unobservable variable U (neither its distribution nor its dimensionality), besides the fact that it influences both $\{X, Y\}$, we are still able to compute the causal effect purely from the observational distribution $P(\boldsymbol{V})$ together with the assumption encoded in $\mathcal{G}$. This together with Eq. 213 allows us to evaluate the effect of any policy $\pi(X)$ from the observational data, i.e.,

$$
\begin{aligned}
\mathbb{E}_\pi[Y] &= \sum_y y \sum_{x,w} P(w|x) \sum_{x'} P(y|w, x')P(x')\pi(x) && (225)\\
&= \sum_{x,w} P(w|x) \sum_{x'} \sum_y yP(y|w, x')P(x')\pi(x) && (226)\\
&= \sum_{x,w} P(w|x) \sum_{x'} \mathbb{E}[Y|w, x']P(x')\pi(x) && (227)
\end{aligned}
$$

This recovers the front-door adjustment formula in Eq. 209. ■



(a) $\mathcal{G}$ (b) $\mathcal{G}_{\overline{X_2}}$ (c) $\mathcal{G}_{\underline{X_2}}$

Figure 22: A causal diagram of the SCM described in Eq. 197 and its manipulated diagrams.

Example 41 *Consider the causal diagram $\mathcal{G}$ in Fig. 22a. We are interested in evaluating the effects of a policy of the form $\pi = (\pi_1(X_1 \mid S_1), \pi_2(X_2 \mid S_1))$. The decomposition in Eq. 214 implies that it is sufficient to identify the interventional distribution $P_{x_2}(Y, S_1)$.*

Our goal then will be to remove the subscript from every probability term in the expression of $P_{x_2}(Y, S_1)$, as shown next. We illustrate the application of the do-calculus in the equations below and with the sub-graphs shown in Figs. 22b-22c. We start by writing the target expression:

$$P_{x_2}(y, s_1) = P_{x_2}(y|s_1)P_{x_2}(s_1) \qquad \textit{Probability Axioms} \tag{228}$$

$$= P_{x_2}(y|s_1)P(s_1) \qquad \textit{Rule 3 } (S_1 \perp\!\!\!\perp X)_{\mathcal{G}_{\overline{X_2}}} \tag{229}$$

$$= P(y|s_1, x_2)P(s_1) \qquad \textit{Rule 2 } (Y \perp\!\!\!\perp X_2 \mid S_1)_{\mathcal{G}_{\underline{X_2}}} \tag{230}$$

The above formula, together with Eq. 214, allows us to identify the expected reward of any policy of the form $\pi = (\pi_1(X_1 \mid S_1), \pi_2(X_2 \mid S_1))$ from the observational data, i.e.,

$$\mathbb{E}_\pi[Y] = \sum_y y \sum_{x_2, s_1} P(y|s_1, x_2)P(s_1)\pi_2(x_2|s_1) \tag{231}$$

$$= \sum_{x_2, s_1} \sum_y yP(y|s_1, x_2)P(s_1)\pi_2(x_2|s_1) \tag{232}$$

$$= \sum_{x_2, s_1} \mathbb{E}[Y|s_1, x_2]P(s_1)\pi_2(x_2|s_1) \tag{233}$$

More specifically, let policy $\pi = (X_1 \leftarrow S_1, X_2 \leftarrow S_1)$. The above equation leads to an evaluation of the expected reward given by:

$$\mathbb{E}_{X_1 \leftarrow S_1, X_2 \leftarrow \neg S_2}[Y] = \sum_{x_2, s_1} \mathbb{E}[Y|s_1, x_2]P(s_1)\mathbb{1}\{x_2 = \neg s_1\} \tag{234}$$

The complete parametrizations for the conditional reward $\mathbb{E}[Y|S_1, X_1]$ and distribution $P(S_1)$ are provided in Table 10. The above equation could thus be further written as:

$$\mathbb{E}_{X_1 \leftarrow S_1, X_2 \leftarrow \neg S_2}[Y] = \mathbb{E}[Y|S_1 = 0, X_2 = 1]P(S_1 = 0) \tag{235}$$

The above computation matches the reward in Eq. 203, as evaluated in SCM $\mathcal{M}^*$. ■

In the above examples, how we apply the rules of do-calculus in the right sequence to obtain a desirable expression is rather unclear. Fortunately, researchers have algorithmatized the procedure to obtain such a sequence, in light of identifiability problems (Tian and Pearl, 2002b; Shpitser and Pearl, 2006b; Huang and Valtorta, 2006a; Bareinboim and Pearl, 2012b; Lee et al., 2019). This means that there exist efficient algorithms to determine the identifiability of the expected rewards of candidate policies from the causal diagram, and if exits, return the identification formula for the target effects from the observational distribution. The algorithms run in a polynomial number of steps relative to the number of nodes and edges in the causal diagram.

S_1	$P(s_1)$
0	0.9
1	0.1

(a) $P(S_1)$

S_1	X_1	$\mathbb{E}[Y\|s_1, x_1]$	S_1	X_1	$\mathbb{E}[Y\|s_1, x_1]$
0	0	0.1	1	0	0.9
0	1	0.9	1	1	0.1

(b) $\mathbb{E}[Y|S_1, X_1]$

Table 10: Evaluation of $P(S_1)$ and $\mathbb{E}[Y|S_1, X_1]$ in SCM $\mathcal{M}^*$ defined in Eq. 197.

4.4 Novel Causal Reinforcement Learning Tasks

We recall the CRL agent is embedded in a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ (Def. 10), where the SCM $\mathcal{M}^*$ is not fully observed, Π represents the policy space, and $\mathcal{R}$ is the reward function. Even though the agent is still evaluated by $\mathcal{M}^*$, we substitute it with the learning regime $\mathcal{L}$ and structural assumptions $\mathcal{A}$ about the environment, which lead to a new signature $\langle \mathcal{L}, \mathcal{A}, \Pi, \mathcal{R} \rangle$, characterizing a *causal reinforcement learning task* (Def. 11). The goal of the agent is then to find a policy π^* such that

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R}\left(\boldsymbol{Y}\right) \mid \ \mathcal{A}, \mathcal{L} \right] \tag{236}$$

In words, the CRL agent aims to find an optimal policy π^* within the policy space Π that maximizes the reward $\mathcal{R}$ when evaluated in the unknown environment $\mathcal{M}^*$ while having assumptions about the environment $\mathcal{A}$ and access to data collected through a learning regime $\mathcal{L}$.

A summary of the signature of the tasks studied so far in this section is shown in the upper part of Table 11, serving as a grounding tool for the discussion here. For instance, each of the tasks accounts for a different dimension in terms of the task signature, as was previously discussed. Off-policy learning considers the more traditional offline modality where the NUC assumption is assumed to hold. In this case, traditional DP or IPW methods could be applied to leverage data collected under one regime – collected under an observational policy – to make inferences about another – a new interventional policy. We also introduced online learning where the learning regime is interventional, and data is collected in an active manner by the agent. Causal assumptions are minimal in this case since the data precisely matches the inferential target. Finally, we studied a more nuanced case of offline learning called "causal identification," which relies on more explicit causal knowledge that allows one to verify the NUC condition or evaluate the optimization given in Eq. 236, even when unconfoundedness doesn't hold. These three modalities touch on different learning regimes and structural assumptions about the underlying $\mathcal{M}^*$.

For the remainder of this paper, we will study natural and pervasive classes of learning tasks that do not fit into these existing modalities but involve novel dimensions and types of analysis relevant to real-world applications. Up next, we list some of these tasks that are also shown in Table 11:

CRL 1. **Causal Offline-to-Online Learning (COOL).** How can we pre-train an online agent to accelerate its learning process by leveraging imperfect knowledge about the effects of candidate policies obtained from confounded observational data?

Computing the effect of candidate policies from observational data might be infeasible, as discussed in Secs. 4.1 and 4.3. On the other hand, it is also undesirable for the AI system to rely solely on brute force, trial-and-error-based experimentation to improve its accuracy. How can we minimize the number of interventions the AI system makes by leveraging the invariances extrapolated from the causal model? In terms of

		Signature				
	Task	**Learning Regime** ($\mathcal{L}$)	**Structural Assumptions** ($\mathcal{A}$)	**Policy Space** (Π)	**Reward Function** ($\mathcal{R}$)	**Section**
1	Off-policy Learning	See	NUC	Π_{EXP}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	4.1
2	Online Learning	Do	-	Π_{EXP}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	4.2
3	Causal Identification	See	DAG $\mathcal{G}$	Π_{EXP}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	4.3
4	Offline-to-Online Learning	See + Do	-	Π_{EXP}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	5
5	Where to do & What to see	Do	DAG $\mathcal{G}$	Π_{MIX}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	6
6	Counterfactual randomization	Ctf-Do	-	Π_{CTF}	$\mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$	7
7	Causal Imitation Learning	See	DAG $\mathcal{G}$	Π_{EXP}	-	8

Table 11: Summary of causal reinforcement learning tasks investigated in this paper, in terms of their signatures and sections. We highlight in gray the most distinct feature introduced by the task.

the prototypical CRL agent depicted in Fig. 11, the green line represents the online learning interactions, while the blue line represents the offline regime using observational data. In Sec. 5, we explore how to combine both modalities when the conditions of offline learning are provably not attainable, yet unlimited experimentation remains undesirable.

CRL 2. **Where to do and What to look for.** Should an agent intervene in the environment to achieve its goal of bringing about a certain state of affairs? If so, where should the intervention take place? The agent's objective is to learn an optimal policy from a collection of candidate policies, each encompassing different actions to intervene and input states to consider when determining these actions, including the null intervention (allowing the system to evolve naturally).

As considered earlier, the agent has a fixed action space Π_{EXP} and tries to identify the intervention $\text{do}(X = x)$ that optimizes its reward measure. In Sec. 6, we explore the structure of the action space in complex systems, Π_{MIX}, focusing on settings where each action plays a qualitatively different role. Practically, this challenge could arise when evaluating the effectiveness of drug combinations given the exponential growth in the total number of possible interactions.

CRL 3. **Counterfactual Decision-Making.** The agent makes a certain decision $X = x$ and wonders: would I be better off had I taken an alternative action, do$(X = x')$? The agent's objective is to evaluate this counterfactual statement to account for its natural and potentially biased decision-making process.

The previous settings considered an experimental policy scope, where typical Fisherian randomization eliminated the agent's natural inclinations. In Sec. 7, we expand the possibilities and allow for such introspective construct, which evokes a new learning regime *Ctf-Do*. This new regime based on what we call counterfactual randomization will allow the agent to navigate through the larger scope of counterfactual policies, Π_{CTF}. In practice, this challenge could appear when evaluating adversarial settings where the agent's natural inclinations were leveraged to trick the agent and minimize its reward in a systematic fashion.

CRL 4. **Causal Imitation Learning.** Does perfectly mimicking an expert always lead to high decision-making performance? If not, under what conditions does imitation learning work? The goal of the agent here is to learn an effective policy from the combination of observational data and a causal diagram when the reward function is not well-specified and unobserved confounding generally exists.

The previous settings we investigated assumed that the reward function was known, which is not always the case. For instance, consider an autonomous vehicle trained from the observed trajectories of a human driver operating the vehicle. It is non-trivial to design a universal reward function evaluating the human's driving performance. How can we program the autonomous vehicle to operate effectively from the demonstration data without knowing the driver's performance measure?

This expanded set with new tasks and understanding paves the way to a broader view of counterfactual learning. It underscores the potential of studying causal inference and reinforcement learning side by side, a program we call *causal reinforcement learning*.

5. Causal Offline-to-Online Learning (CRL Task 1)

Learning algorithms introduced in the previous section rely exclusively on one type of interaction with the underlying environment, either through passive observation ("see"/offline) or direct intervention ("do"/online), despite their strong theoretical guarantees. A natural question that arises is whether the agent could combine both learning regimes and achieve better performance. This leads to the setting of offline-to-online learning. Existing offline-to-online methods in reinforcement learning literature (Taylor and Stone, 2009; Lazaric, 2012; Lee et al., 2022) rely on the NUC assumption (Def. 13), thus are not applicable when unobserved confounders generally exist in the observed data. We will relax the NUC assumption and first study *causal offline-to-online learning* task (for short, COOL) from confounded observational data.[37] This task was studies in (Zhang and

37. More recently, there is growing interest in causal inference to identify treatment effects by combining observational and experimental datasets (Bareinboim and Pearl, 2012b; Lee et al., 2019), and further estimating these under the NUC condition (Colnet et al., 2020; Rosenman et al., 2020; Cho, 2022; Lin and Evans, 2023; Ball et al., 2023), or more general conditions (Jung et al., 2023b,a). These works are orthogonal to offline-to-online learning since they focus on the offline setting where experimental data are provided in priori; the learner does not control the

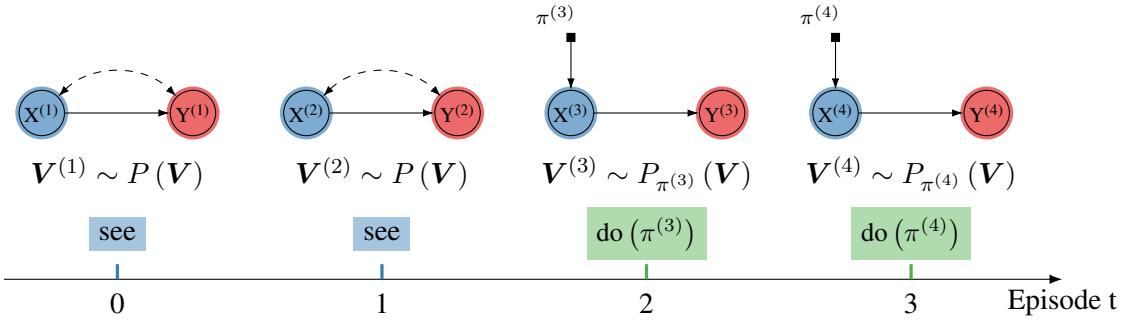


Figure 23: Temporal diagram showing an offline-to-online learning agent interacting with the environment for repeated episodes.

Bareinboim, 2017, 2019), where several algorithms have been proposed. This section will summarize such results under a more unified CRL framework.

The mechanism of how the CRL agent operates in this task and switches from an offline to an online mode when interacting with the underlying environment is illustrated in Fig. 23. Specifically, the CRL agent first passively observes the environment for a number of episodes $t = 1, \dots, n$, and receive the observational samples $\boldsymbol{V}^{(t)} \sim P(\boldsymbol{V})$. For episode $t = n+1, \dots, n+T$, the agent then picks a policy $\pi^{(t)}$, directly intervenes do $\left(\pi^{(t)}\right)$ in the environment, and receives subsequent outcomes $\boldsymbol{V}^{(t)} \sim P_{\pi^{(t)}}(\boldsymbol{V})$. The agent will leverage the observational data $\left\{\boldsymbol{V}^{(1)}, \dots, \boldsymbol{V}^{(n)}\right\}$ to accelerate the future online learning process. The following task signature characterizes this offline-to-online learning setting:

$$\mathcal{T}_{\text{off+on}} = \left\langle \mathcal{I} = \{\text{see}, \text{do}\}, \mathcal{A} = \emptyset, \Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{X_i \in \boldsymbol{X}}, \mathcal{R} = \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R} \right\rangle. \tag{237}$$

This means that the agent will try to find a policy π^* such that

$$\pi^* = \arg\max_{\pi \in \Pi} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R}(\boldsymbol{Y}) \;\middle|\; \mathcal{D}_{\text{obs}} \sim P(\boldsymbol{V}),\ \mathcal{D}_{\text{exp}} \sim P_{\boldsymbol{x}}(\boldsymbol{V}) \right], \tag{238}$$

where the distinct feature here is the combination of observational and interventional interactions.

In order to make this argument more precise, we will describe an online-to-offline strategy that combines the observational data with the learning process of UCB algorithm (Alg. 3), provided that the NUC assumption (Def. 13) holds. Consider an MAB model $\langle \mathcal{M}^*, \{\langle X, \emptyset \rangle\}, Y \rangle$ graphically described in Fig. 10a. Let $\mathcal{D}_{\text{obs}} = \left\{X^{(i)}, Y^{(i)}\right\}_{i=1}^{n}$ be i.i.d. samples drawn from the observational distribution $P(X, Y)$. For UCB algorithm allocating an arm at episode t, let $\mathcal{D}_{\text{exp}}^{(t)} = \left\{X^{(n+i)}, Y^{(n+i)}\right\}_{i=1}^{t-1}$ be the experimental data collected by UCB up to episode t. By combining the observational data $\mathcal{D}_{\text{obs}}$ and experimental data $\mathcal{D}_{\text{exp}}^{(t)}$, we define the empirical reward estimate for every arm x as follows:

$$\hat{\mathbb{E}}_{x}^{(n+t)}[Y] = \frac{1}{N_{n+t}(x)} \left(\sum_{i=1}^{n} Y^{(i)} \mathbb{1}\left\{X^{(i)} = x\right\} + \sum_{i=1}^{t-1} Y^{(n+i)} \mathbb{1}\left\{X^{(n+i)} = x\right\} \right) \tag{239}$$

experiments. On the other hand, the key challenge in COOL is to use observational data to design randomized experiments.

Algorithm 4 Upper Confidence Bound in Direct Offline-to-Online Transfer (UCB$^-$)

1: **Input:** a policy space $\Pi = \{\langle X, \emptyset\rangle\}$, observational data $\mathcal{D}_{\text{obs}} = \left\{X^{(i)}, Y^{(i)}\right\}_{i=1}^{n}$
2: **for all** episodes $t = 1, 2, \ldots$ **do**
3: Choose an arm

$$X^{(n+t)} = \arg\max_{x \in \mathscr{D}(X)} \text{UCB}_{n+t}(x, \delta), \text{ where } \delta = t^{-4} \tag{243}$$

4: Perform do$(X^{(n+t)})$ for episode t and receive reward $Y^{(n+t)}$.
5: **end for**

where $N_{n+t}(x) = \sum_{i=1}^{n+t-1} \mathbb{1}\{X^{(i)} = x\}$ is the total occurrence of observing arm x being played in the combined dataset $\mathcal{D}_{\text{obs}} \cup \mathcal{D}_{\text{exp}}^{(t)}$. The augmented upper confidence bound for an arm x by combining the observational and interventional data is given by

$$\text{UCB}_{n+t}(x, \delta) = \hat{\mathbb{E}}_x^{(n+t)}[Y] + \sqrt{\frac{\log(1/\delta)}{2N_{n+t}(x)}} \tag{240}$$

The augmented UCB algorithm directly transfers observational data as if they were obtained from direct interventions. We summarize in Alg. 4 details of a direct online-to-offline transfer strategy using standard off-policy learning methods, which we call UCB$^-$. For every episode t, it computes an upper confidence bound $\text{UCB}_{n+t}(x, \delta)$ for every arm x, plays an arm with the most significant confidence bound, and observed subsequent reward.

We will analyze the performance of UCB$^-$ and show that the direct transfer strategy could accelerate the UCB's performance, under the NUC assumption (Def. 13). Suppose the NUC condition holds in the MAB model $\langle \mathcal{M}^*, \{\langle X, \emptyset\rangle\}, Y\rangle$. Applying Thm. 3 we compute the expected reward of every arm $x \in \mathscr{D}(X)$ from the observational distribution $P(X, Y)$,

$$\mathbb{E}_x[Y] = \mathbb{E}[Y \mid x] \tag{241}$$

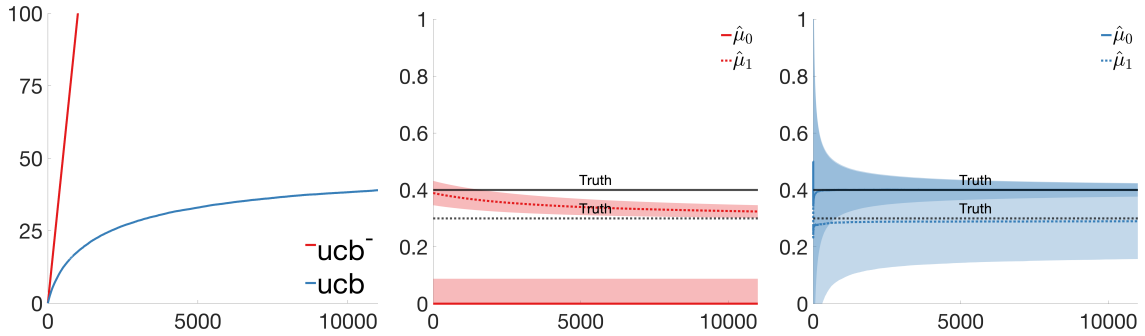
Recall that for any policy $\pi(X)$, the NUC holds in the intervened model $\langle \mathcal{M}_\pi^*, \{\langle X, \emptyset\rangle\}, Y\rangle$. The expected reward of arm x is computable from the interventional distribution $P_\pi(X, Y)$ as

$$\mathbb{E}_x[Y] = \mathbb{E}_\pi[Y \mid x] \tag{242}$$

The above estimation formulas allow the agent to evaluate the effects of arms by pooling the observational and experimental data. Eq. 239 provides consistent estimates for the expected rewards $\mathbb{E}_x[Y]$ provided with the NUC assumption. When sufficient observations are provided, UCB$^-$ will immediately identify the optimal arm; only a few episodes of online interventions are required. Broadly, when the NUC assumption holds, the learner could consistently evaluate candidate policies from the observational data using standard off-policy learning methods. These pre-trained estimations could then be directly transferred to "warm-start" the future online learning process.[38]

On the other hand, the NUC assumption could be fragile and does not hold in many practical applications. For this reason, it is wise to evaluate the performance and robustness of UCB$^-$ when the NUC assumption doesn't hold, which we do in the following experiment.

38. Sec. 4.1 provides a more detailed discussion about the NUC condition and off-policy learning algorithms.



(a) Cumulative Regret (b) UCB- (c) UCB

Figure 24: Simulation results comparing UCB learner with direct transfer of observational data (UCB-) and standard UCB without any prior observations.

Experiment 3 *Fig. 24a shows the cumulative regret of UCB- in the MAB environment $\mathcal{M}^*$ described in Example 1 with the suboptimal gap $\Delta = 0.1$, taking as input $5,000$ observational samples drawn from the distribution $P(X, Y)$. The NUC assumption does not hold in this model due to the unobserved confounder U affecting action X and reward Y simultaneously. As a baseline, we also include a vanilla UCB starting from scratch, which does not utilize any prior observations. One can see by inspection the significant disparity between the performance of UCB (blue) and UCB- (red).*

We show in Fig. 24b the empirical estimates $\hat{\mu}_x$ of the expected rewards $\mathbb{E}_x[Y]$ computed by UCB-; shaded areas represent confidence intervals evaluated at 95% percentile. For comparison, Fig. 24c shows the empirical reward estimates computed by the standard UCB without using prior observations. Simulation results demonstrate a significant bias in the reward estimation of UCB-, favoring the suboptimal arm $x = 1$. This bias was not fully corrected until the end of the online learning process ($T = 10,000$). On the other hand, UCB is able to obtain accurate estimations of the expected rewards after a few episodes of interventions and identify the optimal arm $x^* = 0$. ■

X	$\mathbb{E}[Y\|X = x]$	$\mathbb{E}_{X\leftarrow x}[Y]$	Causal Bound
$x = 0$	0	0.4	$[0, 0.8]$
$x = 1$	$0.5 - 1.25\Delta$	$0.4 - \Delta$	$[0.4 - \Delta, 0.6 - \Delta]$

Table 12: Evaluations of $\mathbb{E}[Y|x]$ and $\mathbb{E}_x[Y]$ in MAB environment $\mathcal{M}^*$ defined in Example 3.

The above example suggests that in MAB models where unobserved confounders exist, the NUC condition does not hold, which implies that directly transferring observational data may introduce a significant bias into the empirical estimation of arms' expected rewards. This, in turn, slows down the learning process of online algorithms, and in some cases may even hinder these algorithms' convergence. In order to confirm this intuition and further explain the negative transfer phenomenon, we compute the expected reward $\mathbb{E}[Y \mid x]$ conditioning on the event that the learner observes arm $X = x$ is played in the MAB environment $\mathcal{M}^*$ defined in Example 3. We also compute the expected reward $\mathbb{E}_x[Y]$ induced by the learner playing an arm $\text{do}(X \leftarrow x)$ in $\mathcal{M}^*$. The analytical results are summarized in Table 12. One can see by inspection that the evaluations of observed expected rewards $\mathbb{E}[Y \mid x]$ differ significantly from the interventional expected rewards $\mathbb{E}_x[Y]$. This means

that the identification formula in Eq. 241 does not apply due to unobserved confounding between the action X and reward Y, which makes some arms x appear observationally more effective than they interventionally are. When the suboptimal gap $\Delta > 0$, optimizing the observed reward $\mathbb{E}[Y \mid x]$ leads to a suboptimal arm $x = 1$. On the other hand, the optimal arm $x^* = 0$ maximizes the interventional reward $\mathbb{E}_x[Y]$ in the underlying MAB environment.

Broadly, off-policy estimation methods may fail to recover the unknown expected rewards of candidate policies without the NUC assumption. Naively transferring estimated rewards introduces inaccuracies in optimal policy estimation of the online learning algorithm, resulting in a negative impact on its performance. Moreover, since the effects of interventions are never measured (before the online learning stage starts), the learner could not detect biases arising from the off-policy evaluation step based on observational data.[39] This implies to significant challenges in offline-to-online learning when the NUC does not generally hold. One may surmise, therefore, that the learner should start the online learning process from scratch without utilizing past observations of the environment, however abundant they are.[40]

This section aims to show that this is not the case and overcome the challenges outlined by the confounded situation as described above. We will study the problem of causal offline-to-online learning (COOL), which accelerates online reinforcement learning by leveraging offline observational data. We focus on the settings where the NUC condition does not hold, and the expected rewards of candidate policies are not computable from the observational data.[41] Directly applying off-policy evaluation could lead to significant bias in the reward estimation, harming the online learning process instead. More specifically, the remainder section is divided as follows.

- Sec. 5.1 introduces a novel causal offline-to-online learning strategy in MAB models and proves that it consistently dominates standard `UCB` algorithm in term of performance. It utilizes bounds to evaluate unknown expected rewards from the observational data, which are then incorporated to accelerate the online learning process.

- Sec. 5.2 generalizes `UCB` algorithm to general CDMs (beyond MABs) where the agent needs to decide on a sequence of actions based on values of corresponding states at the time of intervention. This algorithm achieves the near-optimal regret bound without additional observational data and structural knowledge about the underlying environment.

- Sec. 5.3 derives novel bounds capable of exploiting observational data to infer underlying interventional transitional probabilities and the reward functions. These bounds are then incorporated, in a systematic way, to accelerate online learning in an arbitrary CDM.

5.1 Confounding Robust Offline-to-Online Learning

This section studies the offline-to-online learning in MAB models when the NUC assumption does not hold and standard off-policy learning algorithms, including IPW (Thm. 2) and DP (Thm. 3)

39. On the other hand, whenever the agent goes online, a sufficient test would entail evaluating whether $P_x(Y) = P(Y|X = x)$, which is known as marginal ignorability or no-confounding conditions (Pearl, 2000, Ch. 6); see also (Bareinboim et al., 2020, Def.16(iii)). In practice, the finite-sample version of such test has to be evaluated.

40. This is, of course, against human experience where we learn by observing other agents interacting, even when our perceptions and models of the world do not fully match. Here, we can see that one bit difference between the input of the behavioral agent versus the agent who is using the data can lead to a catastrophic behavior.

41. Of course, whenever NUC holds, this would be a trivial, special case for the approach discussed here.

estimation, do not apply. Causal researchers may wonder if it is possible to estimate the expected rewards of arms from the observational data using causal identification algorithms, e.g., do-calculus learning (Def. 7). Indeed, it has been shown that the expected rewards are not identifiable in MAB environments without additional assumptions. The following corollary could be derived based on the formal definition of identifiability described in Def. 14.

Corollary 3 (Non-Identifiability) *Consider endogenous variables $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{V}$ and let π be a policy over $\boldsymbol{X}$. The interventional (policy) distribution $P_\pi(\boldsymbol{Y})$ is not identifiable from structural assumptions $\mathcal{A}$ and observational distribution $P(\boldsymbol{V}$ if there exist two SCMs $\mathcal{M}_1, \mathcal{M}_2$ compatible with $\mathcal{A}$ such that $P(\boldsymbol{V}; \mathcal{M}_1) = P(\boldsymbol{V}; \mathcal{M}_2) > 0$ while $P_\pi(\boldsymbol{Y}; \mathcal{M}_1) \neq P_\pi(\boldsymbol{Y}; \mathcal{M}_2)$.* ■

In words, the expected rewards $\mathbb{E}_x[Y]$ of arms x are not identifiable in MAB models if there exist two MAB environments that generate the same observational distribution $P(X, Y)$, but differ in the expected rewards $\mathbb{E}_x[Y]$. This means the agent could not uniquely determine the expected rewards of arms from the observational distribution alone. Our next example demonstrates this non-identifiability result in MAB models.

Example 42 *Consider an MAB environment $\mathcal{M}'$ described by an SCM*

$$\mathcal{M}' = \langle \boldsymbol{U} = \{U_1, U_2\}, \boldsymbol{V} = \{X, Y\}, \mathscr{F}', P(U_1, U_2) \rangle, \tag{244}$$

The causal mechanisms are the following:

$$\mathscr{F}' = \begin{cases} X \leftarrow \mathbb{1}\{U_1 < 0.8\}, \\ Y \leftarrow \mathbb{1}\{U_2 < 0.5 - 1.25\Delta\} \times X \end{cases} \tag{245}$$

where coefficient Δ is a real number bounded in $(0, 0.5)$; and $P(U_1, U_2)$ is such that U_1, U_2 are independent variables drawn from a uniform distribution $\texttt{Unif}(0, 1)$. It is verifiable that $\mathcal{M}'_{\text{MAB}}$ defines the same observational distribution $P(X, Y)$ as the MAB environment $\mathcal{M}^*$ defined in Example 3. First, marginal probabilities $P(X = 0) = 0.2$ and $P(X = 1) = 0.8$ in $\mathcal{M}'_{\text{MAB}}$ since U_1 is uniformly drawn from the real interval $[0, 1]$. Evaluating the conditional distribution $P(Y|X)$ in $\mathcal{M}'$ gives

$$\begin{aligned} P(Y = 1 \mid X = 0) &= P(\mathbb{1}\{U_2 < 0.5 - 1.25\Delta\} \times 0 = 1 \mid X = 0) \\ &= 0 \end{aligned} \tag{246}$$

Similarly, the recovery rate conditioning on event $X = 1$ is given by

$$\begin{aligned} P(Y = 1 \mid X = 1) &= P(U_2 < 0.5 - 1.25\Delta) \\ &= 0.5 - 1.25\Delta. \end{aligned} \tag{247}$$

On the other hand, $\mathcal{M}'$ defines different expected rewards for interventions $do(X \leftarrow x)$ from that defined by $\mathcal{M}'$. More precisely, the submodel $\mathcal{M}_x$ induced by $do(X \leftarrow x)$ is a tuple

$$\mathcal{M}'_x = \langle \boldsymbol{U} = \{U_1, U_2\}, \boldsymbol{V} = \{X, Y\}, \mathscr{F}'_x, P(U_1, U_2) \rangle, \tag{248}$$

where the structural functions $\mathscr{F}'_x$ is defined as

$$\mathscr{F}'_x = \begin{cases} X \leftarrow x, \\ Y \leftarrow \mathbb{1}\{U_2 < 0.5 - 1.25\Delta\} \times X \end{cases} \tag{249}$$

Evaluating the expected reward Y in submodel $\mathcal{M}'_{X\leftarrow 0}$ described in Eq. 249 gives

$$\begin{aligned}\mathbb{E}_{X\leftarrow 0}[Y] &= \mathbb{E}[\mathbb{1}\{U_2 < 0.5 - 1.25\Delta\} \times 0] \\ &= 0\end{aligned} \tag{250}$$

Similarly, the expected reward of playing an arm do$(X \leftarrow 1)$ is equal to

$$\begin{aligned}\mathbb{E}_{X\leftarrow 1}[Y] &= P(U_2 < 0.5 - 1.25\Delta) \\ &= 0.5 - 1.25\Delta.\end{aligned} \tag{251}$$

For detailed computations of $P(X, Y)$ and $\mathbb{E}_x[Y]$ in MAB model $\mathcal{M}^*$, revisit Examples 3 and 7. To sum up, MAB models $\mathcal{M}'$ and $\mathcal{M}^*$ are both compatible with the causal diagram $\mathcal{G}_{\text{MAB}}$ of Fig. 10a such that they define the same observational distribution $P(X, Y)$ while differ significantly in the expected reward $\mathbb{E}_x[Y]$. This means the expected rewards of arms x are not identifiable from the observational distribution $P(X, Y)$ in MAB models. ■

The result so far seems to suggest that when unobserved confounders exist and no additional causal knowledge is provided, no prior observations could be useful in evaluating the expected rewards of arms in MAB models. However, we will show this is not the case by deriving bounds over the unknown expected rewards from the observational data, which we call *causal bounds*. This means that while it is infeasible to determine the values of the non-identifiable expected reward, the learner could still extrapolate partial knowledge from the observational data to improve the estimates of its feasible region. For MAB models with an arm choice X and a reward Y, the seminal results in (Manski, 1990; Robins, 1989) allow the derivation of informative causal bounds (to be defined) containing the expected reward of an arm x from the observational distribution.

Theorem 8 (Natural Bounds (Manski, 1990)) *For any SCM $\mathcal{M}^*$ containing an action X and a reward Y, let the domain of X be discrete and finite, and Y be bounded in the real interval $[0, 1]$. The expected reward for any arm x is bounded in $\mathbb{E}_x[Y] \in [l_x, r_x]$ where*

$$l_x = \underbrace{\mathbb{E}[Y \mid x]\, P(x)}_{observational}, \qquad r_x = \underbrace{\mathbb{E}[Y \mid x]\, P(x) + 1 - P(x)}_{observational} \tag{252}$$

■

The lower and upper bounds in Eq. 252 are both functions of the observational distribution $P(X, Y)$ and are, therefore, estimable from the observational data. The above bounds are informative and strictly contained in the interval $[0, 1]$ when marginal probabilities $P(x) > 0$ are positive for any arm $x \in \mathscr{D}(X)$. The natural bounds have been proved to be optimal in MAB models (Zhang and Bareinboim, 2017, 2021), i.e., they cannot be improved without additional assumptions and data.

Example 43 *Consider the MAB environments $\mathcal{M}^*, \mathcal{M}'$ described in Examples 1 and 42, respectively. Applying Thm. 8 we obtain a lower bound contained the expected reward $\mathbb{E}_{X\leftarrow 0}[Y]$ computed from the observational distribution $P(X, Y)$ as*

$$\begin{aligned}\mathbb{E}_{X\leftarrow 0}[Y] &\geq \mathbb{E}[Y \mid X = 0]\, P(X = 0) \\ &= 0.\end{aligned} \tag{253}$$

The upper bound over the expected reward for arm $x = 0$ is given by:

$$\begin{aligned}\mathbb{E}_{X\leftarrow 0}[Y] &\leq \mathbb{E}[Y \mid X = 0]\, P(X = 0) + 1 - P(X = 0) \\ &\leq 0.8 \end{aligned} \tag{254}$$

Similarly, we could also obtain a natural bound for the expected reward of arm $x = 1$ from the observational distribution $P(X, Y)$ and is given by

$$\begin{aligned}\mathbb{E}_{X\leftarrow 1}[Y] &\geq \mathbb{E}[Y \mid X = 1]\, P(X = 1) \\ &= (0.5 - 1.25\Delta) \times 0.8 \\ &= 0.4 - \Delta. \end{aligned} \tag{255}$$

and the upper bound implies

$$\begin{aligned}\mathbb{E}_{X\leftarrow 1}[Y] &\leq \mathbb{E}[Y \mid X = 1]\, P(X = 1) + 1 - P(X = 1) \\ &\leq 0.4 - \Delta + P(X = 0) \\ &\leq 0.6 - \Delta \end{aligned} \tag{256}$$

We summarize in Table 12 natural bounds computed from the observational distribution $P(X, Y)$. The results support the soundness of the natural bound in Thm. 8 in MAB models since it contains the real expected rewards $\mathbb{E}_x[Y]$ evaluated in both $\mathcal{M}^*$ and $\mathcal{M}'$. ■

The causal bounds $\mathbb{E}_x[Y] \in [l_x, r_x]$ could be used to improve the estimate of the upper confidence bound assigned to every arm x during online learning. More precisely, for UCB algorithm selecting an arm at episode t, the causally-clipped upper confidence bound for arm x is defined as

$$\overline{\text{UCB}}_t(x, \delta) = \min\{\max\{\text{UCB}_t(x, \delta), l_x\}, r_x\}. \tag{257}$$

Among quantities in the above equation, $\text{UCB}_t(x, \delta)$ is the standard upper confidence bound for MAB models defined in Eq. 182. The clipping ensures the new upper bound $\overline{\text{UCB}}_t(x, \delta) \in [l_x, r_x]$ is contained in the causal bound. In words, the causal bound for an arm x always takes priority when it is incompatible with the confidence bound computed from the experimental data collected from online learning to episode t. The incompatibilities generally arise at the beginning of the learning process (t is small) where the standard upper bound $\text{UCB}_t(x, \delta)$ is loose. It eventually converges between the causal bound $[l_x, r_x]$ as more experimental data $N_t(x)$ are collected. The algorithm consistently prefers causal bounds since the observational data is often abundant while conducting interventions is expensive; causal bounds in Thm. 8 are valid and could be accurately estimated with sufficient observational data.

Alg. 5 summarizes the augmented UCB algorithm incorporating causal bounds computed from the observational data, which we call UCB^+. It takes as input arguments causal bounds $\mathbb{E}_x[Y] \in [l_x, r_x]$ over the expected rewards for candidate arms x. For every episode t, it computes the clipped upper bound $\overline{\text{UCB}}_{n+t}(x, \delta)$ for every arm x by combining the experimental data collected up to episode t and the causal bound $[l_x, r_x]$. Finally, the agent plays an arm with the highest clipped confidence bound and receives a subsequent reward.

Algorithm 5 Upper Confidence Bound combined with Causal Bounds in MAB (UCB$^+$)

1: **Input:** a policy space $\Pi = \{\langle X, \emptyset \rangle\}$, causal bounds $\mathbb{E}_x[Y] \in [l_x, r_x]$.
2: **for all** episodes $t = 1, 2, \ldots$ **do**
3: Choose an arm

$$X^{(t)} = \arg\max_{x \in \mathscr{D}(X)} \overline{\text{UCB}}_{n+t}(x, \delta), \text{ where } \delta = t^{-4}. \tag{258}$$

4: Perform do$(X^{(t)})$ for episode t and receive reward $Y^{(t)}$.
5: **end for**

Theorem 9 *For any MAB model $\langle \mathcal{M}^*, \{\langle X, \emptyset \rangle\}, Y \rangle$, let Y be the reward variable with support on $[0, 1]$ and let the domain of action X be $\mathscr{D}(X) = \{1, \ldots, K\}$. It holds the regret of UCB$^+$ in SCM $\mathcal{M}^*$ after $T > 1$ episodes is bounded by*

$$R(T, \mathcal{M}^*) \leq 8 \sum_{\substack{x: \Delta_x > 0 \\ r_x \geq \mu_{x^*}}} \frac{\log(T)}{\Delta_x} + \left(1 + \frac{\pi^2}{3}\right) \sum_{x: \Delta_x > 0} \Delta_x \tag{259}$$

■

Let $\mathscr{D}(X)^- = \{x \in \mathscr{D}(X) \mid \Delta_x > 0\}$ be the set of suboptimal arms. Let $\mathscr{D}(X)^*$ be the set of suboptimal arms such that the causal upper bound r_x for every arm x is larger than or equal to the optimal expected reward $\mu_{x^*} = \mathbb{E}_{x^*}[Y]$, i.e., $\mathscr{D}(X)^* = \{x \in \mathscr{D}(X) \mid \Delta_x > 0, r_x \geq \mu^*\}$. Since $\mathscr{D}(X)^* \subseteq \mathscr{D}(X)^-$, the regret bound of Thm. 9 consistently dominates the regret bound of the standard UCB in Thm. 5. When there are some suboptimal arms x with $r_x < \mu_{x^*}$, the augmented UCB$^+$ is able to outperform UCB by utilizing quantitative knowledge extrapolated from the observational data.[42]

To illustrate, assume the total number of arms $K = 2$ and $x = 0$ is the optimal arm; the suboptimal gap $\Delta = \mathbb{E}_{X \leftarrow 0}[Y] - \mathbb{E}_{X \leftarrow 1}[Y]$. Applying UCB gives the regret bound

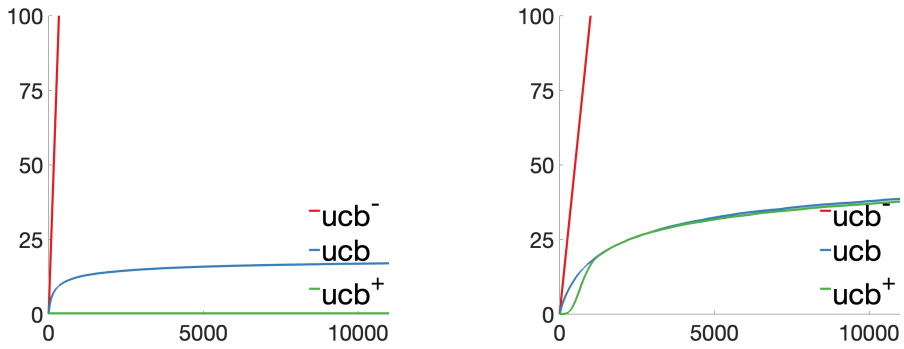
$$R(T, \mathcal{M}^*) \leq \frac{8 \log(T)}{\Delta} + \left(1 + \frac{\pi^2}{3}\right) \Delta \tag{260}$$

Suppose that the causal bound $[l_1, r_1]$ of arm $x = 1$ is informative and $r_1 < \mu_0$. Thm. 9 implies that the regret bound of UCB$^+$ taking into account this causal bound is

$$R(T, \mathcal{M}^*) \leq \left(1 + \frac{\pi^2}{3}\right) \Delta \tag{261}$$

In words, when the causal bound is informative, UCB$^+$ enjoys a constant regret $\mathcal{O}(1)$ which is orders of magnitude smaller than the sublinear regret $\mathcal{O}(\log(T)/\Delta)$ of UCB. On the other hand, if the causal bound is not informative and $r_1 \geq \mu_0$, the regret bound of UCB$^+$ coincides with the regret of UCB in Eq. 260, and no negative transfer occurs.

42. Another line of popular bandit algorithm is called Thompson sampling (TS, Thompson (1933); Chapelle and Li (2011); Agrawal and Goyal (2012)). We show in Appendix C that causal bounds could also be utilized to accelerate the convergence of the TS algorithm.



(a) $\Delta = 0.3$ (b) $\Delta = 0.1$

Figure 25: Simulation results comparing $\texttt{UCB}^{+}$ learner augmented with causal bounds over the expected rewards, standard $\texttt{UCB}$, and $\texttt{UCB}^{-}$ with direct transfer of observational data.

Experiment 4 *Fig. 25 shows the cumulative regret of* $\texttt{UCB}^{+}$ *in the MAB environment* $\mathcal{M}^*$ *described in Example 1 with the suboptimal gap* $\Delta = 0.3$ *and* $\Delta = 0.1$ *respectively. It takes as input the natural bounds* $[l_x, r_x]$ *over the expected rewards of arm* $x = 0, 1$*, computed from taking as input* $5,000$ *observational samples drawn from the distribution* $P(X, Y)$*. As a baseline, we also include a vanilla* $\texttt{UCB}$ *starting from scratch without utilizing any prior observations, and* $\texttt{UCB}^{-}$ *with the confounded observational data directly transferred.*

One can see by inspection the significant disparity between the performance of $\texttt{UCB}^{+}$ *and* $\texttt{UCB}$ *for* $\Delta = 0.3$*. In this case, the causal bound for the suboptimal arm* $x = 1$ *is* $r_1 = 0.6 - \Delta = 0.3 < \mu_0$*, and* $\texttt{UCB}^{+}$ *converges to the optimal arm* $x = 0$ *almost immediately after the learning starts. On the other hand, when the suboptimal gap* $\Delta = 0.1$ *and the causal bound* $r_1 = 0.6 - \Delta = 0.5 > \mu_0$*, the performance of* $\texttt{UCB}^{+}$ *and* $\texttt{UCB}$ *virtually coincides. The simulation results corroborate the theory that the transfer strategy of* $\texttt{UCB}^{+}$ *enjoys no negative impact. As expected, the direct transfer* $\texttt{UCB}^{-}$ *performs the worst among all strategies due to unobserved confounding.* ■

5.2 Online Learning in Sequential Decision-Making

The offline-to-online learning strategy described so far focuses on the MAB models with a single decision horizon $H = 1$. The remainder of this section will extend this strategy to optimize a general CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ where the policy space $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}$ and the reward function $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$ taking a set of signals $\boldsymbol{Y}$ as input. We will first introduce a purely online learning algorithm to optimize a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ without detailed parametrization of the underlying environment $\mathcal{M}^*$. Compared to bandit algorithms previously described for MAB models, our proposed algorithm also interacts with the underlying SCM $\mathcal{M}^*$ for repeated episodes $t = 1, \ldots, T$. For each episode t, instead of selecting a single arm X, our algorithm will determine values of a sequence of actions $X_1, \ldots, X_H$. More specifically, it will pick a policy $\pi^{(t)} = \left(\pi_1^{(t)}, \ldots, \pi_H^{(t)}\right)$ at the beginning of episode t. For every step $i = 1, \ldots, H$ of interventions, our algorithm observed state $\boldsymbol{S}_i^{(t)}$, selects an action $X_i^{(t)} \sim \pi_i^{(t)}\left(X_i \mid \boldsymbol{S}_i^{(t)}\right)$ following the decision rule $\pi_i^{(t)}$, and perform intervention do $\left(X_i \leftarrow X_i^{(t)}\right)$. The cumulative regret for an online learning algorithm operating in a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ after $T > 1$ episodes of interventions is defined as follows, compared to an

idealized agent following the optimal policy π^* for all episodes of interactions $t = 1, \ldots, T$,

$$R(T, \mathcal{M}^*) = T\mathbb{E}_{\pi^*}\left[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*\right] - \sum_{t=1}^{T} \mathbb{E}_{\pi^{(t)}}\left[\mathcal{R}\left(\boldsymbol{Y}\right)\right]. \tag{262}$$

Similar to the bandit setting, a nice property for an online algorithm is to achieve a sublinear regret $R(T, \mathcal{M}^*) = o(T)$ so that it eventually converges to an optimal policy π^*.[43]

There exists an experimental design of randomized trials for optimizing policies over a finite sequence of actions $\boldsymbol{X}$ in an unknown environment, called sequential multiple assignment randomized trials (for short, SMART (Murphy, 2005a)). It is an explore-then-commit algorithm. More specifically, fix a total number of trials $N \in \mathbb{N}$. For the first $t \leq N$ episodes, the SMART algorithm explores by sampling values $X_i^{(t)}$ of every action X_i, for $i = 1, \ldots, H$, from its domain $\mathscr{D}(X_i)$ uniformly at random. For episodes $t > N$, the algorithm commits to a policy $\pi^{(t)}$ maximizing the empirical reward estimates $\hat{\mathbb{E}}_{\pi}\left[\mathcal{R}(\boldsymbol{Y})\right]$ computed from experimental data $\mathcal{D}_{\text{exp}}^{(N)}$ collected during exploration. Details of the algorithm is summarized in Alg. 2.

When the total number of trials N is sufficiently large, SMART is able to recover the effects of candidate policies and find an optimal policy from the experimental data (Murphy et al., 2001a; Murphy, 2005b). However, as previously discussed in Sec. 4.2, explore-then-commit algorithms suffer from a linear regret during the exploration phase ($t \leq N$). Determining the optimal trial number N is theoretically challenging, requiring prior parametric knowledge about the underlying environment and the total episodes of interactions T.

We will next describe a novel online learning algorithm for optimizing a policy space over a sequence of actions $\boldsymbol{X} = \{X_1, \ldots, X_H\}$. It is able to achieve a sublinear regret $R(T, \mathcal{M}^*) = o(T)$ without parametric knowledge of the SCM $\mathcal{M}^*$ and total episodes T. Our discussion begins with the decision with some necessary notations and technical tools. Recall that for every $i = 1, \ldots, H$, $\bar{\boldsymbol{X}}_i = \{X_1, \ldots, X_i\}$ is a sequence of actions up to stage i and $\bar{\boldsymbol{S}}_i = \{\boldsymbol{S}_1, \ldots, \boldsymbol{S}_i\}$ is the sequence of corresponding states. For any policy $\pi \in \Pi$, using basic probabilistic operations and the Bayes' rule, the expected reward $\mathbb{E}_{\pi}\left[\mathcal{R}(\boldsymbol{Y})\right]$ could be written as

$$\mathbb{E}_{\pi}\left[\mathcal{R}(\boldsymbol{Y})\right] = \sum_{\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H} \underbrace{\mathbb{E}_{\pi}\left[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H\right]}_{\text{reward}} \prod_{i=0}^{H-1} \underbrace{P_{\pi}\left(s_{i+1} \mid \bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i\right)}_{\text{transition probabilities}} \underbrace{\pi_{i+1}\left(x_{i+1} \mid \boldsymbol{s}_{i+1}\right)}_{\text{policy}} \tag{263}$$

The above equation follows that in submodel $\mathcal{M}_{\pi}^*$, values of every action X_i are determined by the function π_i. Among the quantities above, probabilities of policy π are known. It is sufficient to estimate transition distributions $P_{\pi}\left(s_{i+1} \mid \bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i\right)$ and the conditional reward $\mathbb{E}_{\pi}\left[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H\right]$.

The NUC condition holds in the post-interventional system $\langle \mathcal{M}_{\pi}^*, \Pi, \mathcal{R} \rangle$ (Lem. 1). Therefore, for every $i = 0, \ldots, H-1$, conditioning on past states $\bar{\boldsymbol{S}}_i$ d-separates all backdoor paths between actions $\bar{\boldsymbol{X}}_i$ and another variables in submodel $\mathcal{M}_{\pi}^*$. Applying Rule 2 of do-calculus (Thm. 7),

$$P_{\pi}\left(s_{i+1} \mid \bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i\right) = P_{\bar{\boldsymbol{x}}_i}\left(s_{i+1} \mid \bar{\boldsymbol{s}}_i\right), \tag{264}$$

$$\mathbb{E}_{\pi}\left[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H\right] = \mathbb{E}_{\bar{\boldsymbol{x}}_H}\left[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{s}}_H\right] \tag{265}$$

In words, the transition distribution $P_{\pi}\left(s_{i+1} \mid \bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i\right)$ and conditional reward $\mathbb{E}_{\pi}\left[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H\right]$ remain invariant across policies $\pi \in \Pi$. Therefore, they could be consistently estimated by pooling interventional data collected by different candidate policies in Π.

43. See Sec. 4.2 for a detailed discussion of the online learning task and properties of cumulative regret.

Algorithm 6 Upper Confidence Bound (UCB) for CDMs

Require: a policy space $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}$, a reward function $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto [0, 1]$

1: **for all** episodes $t = 1, 2, \ldots$ **do**

2: Construct a set $\mathbb{M}_{t-1}(\delta)$ of all candidate SCMs $\mathcal{M}$, with error probability $\delta = t^{-4}$, that are compatible with the interventional data $\left\{\boldsymbol{V}^{(1)}, \ldots, \boldsymbol{V}^{(t-1)}\right\}$ collected up to episode t.

3: Find the optimal policy $\pi^{(t)}$ of an optimistic SCM $\mathcal{M}^{(t)} \in \mathbb{M}_{t-1}(\delta)$ such that

$$\mathbb{E}_{\pi^{(t)}}\left[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^{(t)}\right] = \max_{\pi, \mathcal{M}} \mathbb{E}_{\pi}\left[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}\right] \text{ s.t. } \pi \in \Pi, \mathcal{M} \in \mathbb{M}_{t-1}(\delta). \tag{269}$$

4: Perform do $\left(\pi^{(t)}\right)$ for episode t and receive observations $\boldsymbol{V}^{(t)}$.

5: **end for**

Throughout this section, we will consistently assume that the domains of the states $\boldsymbol{S}$ and actions $\boldsymbol{X}$ are finite; the reward function $\mathcal{R}(\boldsymbol{Y})$ is bounded in a real interval $[0, 1]$. Fix a finite sequence $\pi^{(1)}, \ldots, \pi^{(t)} \in \Pi$. Given finite samples $\left\{\boldsymbol{V}^{(1)}, \ldots, \boldsymbol{V}^{(t)}\right\}$ drawn from distributions $P_{\pi^{(1)}}(\boldsymbol{V}) \ldots, P_{\pi^{(t)}}(\boldsymbol{V})$ respectively, empirical mean estimates for the transition distribution $P_{\bar{\boldsymbol{x}}_{i-1}}(\boldsymbol{S}_i \mid \bar{\boldsymbol{s}}_{i-1})$, $i = 1, \ldots, H-1$, and the conditional reward $\mathbb{E}_{\bar{\boldsymbol{x}}_H}[Y \mid \bar{\boldsymbol{s}}_H]$ are defined as:

$$\forall i = 1, \ldots, H-1, \ \hat{P}^{(t)}_{\bar{\boldsymbol{x}}_i}(\boldsymbol{s}_{i+1} \mid \bar{\boldsymbol{s}}_i) = \frac{\sum_{j=1}^{t} \mathbb{1}\left\{\bar{\boldsymbol{X}}_i^{(j)} = \bar{\boldsymbol{x}}_i, \bar{\boldsymbol{S}}_{i+1}^{(j)} = \bar{\boldsymbol{s}}_{i+1}\right\}}{N_t(\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i)}, \tag{266}$$

$$\text{and } \hat{\mathbb{E}}^{(t)}_{\bar{\boldsymbol{x}}_H}[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{s}}_H] = \frac{\sum_{j=1}^{t} \mathbb{1}\left\{\bar{\boldsymbol{X}}_H^{(j)} = \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{S}}_H^{(j)} = \bar{\boldsymbol{s}}_H\right\} \mathcal{R}\left(\boldsymbol{Y}^{(j)}\right)}{N_t(\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H)}, \tag{267}$$

Among quantities in the above equations, for every $i = 1, \ldots, H$, $N_t(\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i)$ is the event count for every state-action pair $(\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i) \in \mathscr{D}(\bar{\boldsymbol{X}}_i \cup \bar{\boldsymbol{S}}_i)$ defined as

$$\forall i = 1, \ldots, H, \ \ N_t(\bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i) = \max\left\{1, \sum_{j=1}^{t} \mathbb{1}\left\{\bar{\boldsymbol{X}}_i^{(j)} = \bar{\boldsymbol{x}}_i, \bar{\boldsymbol{S}}_i^{(j)} = \bar{\boldsymbol{s}}_i\right\}\right\}. \tag{268}$$

Alg. 6 shows details of UCB algorithm capable of optimizing an unknown CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$. It works in phases of model construction, optimistic planning, and policy execution. In Step 2, UCB constructs a set $\mathbb{M}_{t-1}(\delta)$ of plausible SCMs from interventional data $\left\{\boldsymbol{V}^{(1)}, \ldots, \boldsymbol{V}^{(t-1)}\right\}$. For every SCM $\mathcal{M} \in \mathbb{M}_{t-1}(\delta)$, its transition distributions $P_{\bar{\boldsymbol{x}}_i}\left(\boldsymbol{S}_{i+1} \mid \bar{\boldsymbol{S}}_i\right)$, $i = 1, \ldots, H-1$, and the conditional reward $\mathbb{E}_{\bar{\boldsymbol{x}}_H}[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{s}}_H]$ are contained in convex intervals centering around their corresponding empirical estimates computed from interventional data collected prior to episode t. The error probability δ is set as a decreasing function of the episode number t so that $\mathbb{M}_{t-1}(\delta)$ contains the underlying SCM $\mathcal{M}^*$ with high probability as the algorithm continues and more interventional data are collected. It then computes in Step 3 an optimal policy $\pi^{(t)}$ of the most optimistic SCM $\mathcal{M}^{(t)} \in \mathbb{M}_{t-1}(\delta)$ that induces the maximal expected reward. We will discuss the details of the model construction and the optimistic planning later. Finally, policy $\pi^{(t)}$ is executed throughout episode t and new samples $\boldsymbol{V}^{(t)} \sim P_{\pi^{(t)}}(\boldsymbol{V})$ are collected (Step 4).

Model Construction Let $\left\{Y^{(1)}, \ldots, Y^{(n)}\right\}$ be i.i.d. samples drawn from a discrete distribution $P(Y)$. Let $\hat{P}(y) = \frac{1}{n}\sum_{t=1}^{n} \mathbb{1}\left\{Y^{(i)} = y\right\}$ be empirical estimates of probabilities $P(y)$. Generally,

the L_1-deviation of the true distribution and the empirical distribution is bounded according to (Weissman et al., 2003)

$$P\left(\left\|\hat{P}\left(\cdot\right)-P\left(\cdot\right)\right\|_1>\sqrt{\frac{2\left|\mathscr{D}(Y)\right|\log\left(2/\delta\right)}{n}}\right)\leq\delta \tag{270}$$

Fix $\delta\in(0,1)$. Let $\mathbb{M}_t(\delta)$ be the set of all SCMs $\mathcal{M}$ with endogenous variables $\boldsymbol{V}$, and with its transition distribution $P_{\bar{\boldsymbol{x}}_i}\left(\boldsymbol{S}_{i+1}\mid\bar{\boldsymbol{s}}_i\right)$ close to the empirical distribution $\hat{P}^{(t)}_{\bar{\boldsymbol{x}}_i}\left(\boldsymbol{S}_{i+1}\mid\bar{\boldsymbol{s}}_i\right)$, $i=0,\ldots,H-1$, and the reward $\mathbb{E}_{\bar{\boldsymbol{x}}_H}\left[\mathcal{R}(\boldsymbol{Y})\mid\bar{\boldsymbol{s}}_H\right]$ close to the empirical reward $\hat{\mathbb{E}}^{(t)}_{\bar{\boldsymbol{x}}_H}\left[\mathcal{R}(\boldsymbol{Y})\mid\bar{\boldsymbol{s}}_H\right]$, i.e.,

$$\forall i=0\ldots,H-1,\quad\left\|P_{\bar{\boldsymbol{x}}_i}(\cdot|\bar{\boldsymbol{s}}_i;\mathcal{M})-\hat{P}^{(t)}_{\bar{\boldsymbol{x}}_i}(\cdot|\bar{\boldsymbol{s}}_i)\right\|_1\leq\epsilon_i(\delta), \tag{271}$$

$$\text{and}\quad\left|\mathbb{E}_{\bar{\boldsymbol{x}}_H}[\mathcal{R}(\boldsymbol{Y})|\bar{\boldsymbol{s}}_H;\mathcal{M}]-\hat{\mathbb{E}}^{(t)}_{\bar{\boldsymbol{x}}_H}[\mathcal{R}(\boldsymbol{Y})|\bar{\boldsymbol{s}}_H]\right|\leq\epsilon_H(\delta). \tag{272}$$

where the confidence width $\epsilon_i(\delta)$ is a function given by,

$$\forall i=0,\ldots,H-1,\ \epsilon_i(\delta)=\sqrt{\frac{2\left|\mathscr{D}(\boldsymbol{S}_{i+1})\right|\log(4H\left|\mathscr{D}(\bar{\boldsymbol{S}}_i\cup\bar{\boldsymbol{X}}_i)\right|/\delta)}{N_t(\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{s}}_i)}} \tag{273}$$

$$\text{and}\quad\epsilon_H(\delta)=\sqrt{\frac{\log(8H\left|\mathscr{D}(\bar{\boldsymbol{S}}_H\cup\bar{\boldsymbol{X}}_H)\right|/\delta)}{2N_t(\bar{\boldsymbol{x}}_H,\bar{\boldsymbol{s}}_H)}} \tag{274}$$

Applying a union bound over the concentration inequalities in Eq. 270 and Hoeffding's inequality in Eq. 181, we obtain the following error probability:

$$P\left(\mathcal{M}^*\notin\mathbb{M}_t(\delta)\right)\leq\sum_{i=0}^{H-1}\sum_{\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{s}}_i}P\left(\left\|P_{\bar{\boldsymbol{x}}_i}(\cdot|\bar{\boldsymbol{s}}_i;\mathcal{M}^*)-\hat{P}^{(t)}_{\bar{\boldsymbol{x}}_i}(\cdot|\bar{\boldsymbol{s}}_i)\right\|_1>\epsilon_i(\delta)\right) \tag{275}$$

$$+\sum_{\bar{\boldsymbol{x}}_H,\bar{\boldsymbol{s}}_H}P\left(\left|\mathbb{E}_{\bar{\boldsymbol{x}}_H}[\mathcal{R}(\boldsymbol{Y})|\bar{\boldsymbol{s}}_H;\mathcal{M}]-\hat{\mathbb{E}}^{(t)}_{\bar{\boldsymbol{x}}_H}[\mathcal{R}(\boldsymbol{Y})|\bar{\boldsymbol{s}}_H]\right|>\epsilon_H(\delta)\right) \tag{276}$$

$$=\sum_{i=0}^{H-1}\sum_{\bar{\boldsymbol{x}}_i,\bar{\boldsymbol{s}}_i}\frac{\delta}{2H\left|\mathscr{D}(\bar{\boldsymbol{S}}_i\cup\bar{\boldsymbol{X}}_i)\right|}+\sum_{\bar{\boldsymbol{x}}_H,\bar{\boldsymbol{s}}_H}\frac{\delta}{2H\left|\mathscr{D}(\bar{\boldsymbol{S}}_H\cup\bar{\boldsymbol{X}}_H)\right|} \tag{277}$$

$$<\delta \tag{278}$$

That is, the underlying SCM $\mathcal{M}^*$ is contained in SCM family $\mathbb{M}_t(\delta)$ with probability at least $1-\delta$.

Optimistic Planning Step 7 of `UCB` tries to find an optimal policy $\pi^{(t)}$ for an optimistic SCM $\mathcal{M}^{(t)}$. For a fixed SCM $\mathcal{M}$, standard planning algorithms (Bellman, 1957; Koller and Milch, 2003) are applicable to allow one to find an optimal policy in space Π that maximizes the expected reward. However, the optimization problem of Eq. 269 also requires the learner to find an SCM $\mathcal{M}^{(t)}$ that defines the maximal optimal reward among all plausible SCMs in family $\mathbb{M}_{t-1}(\delta)$.

Generally, we can formulate this problem as a polynomial program as follows. The decomposition in Eq. 263, together the invariances in Eqs. 264 and 265, allows one to write the expected reward $\mathbb{E}_\pi\left[\mathcal{R}(\boldsymbol{Y})\right]$ as follows

$$\mathbb{E}_\pi\left[\mathcal{R}(\boldsymbol{Y})\right]=\sum_{\bar{\boldsymbol{x}}_H,\bar{\boldsymbol{s}}_H}\mathbb{E}_{\bar{\boldsymbol{x}}_H}\left[\mathcal{R}(\boldsymbol{Y})\mid\bar{\boldsymbol{s}}_H\right]\prod_{i=0}^{H-1}P_{\bar{\boldsymbol{x}}_i}\left(s_{i+1}\mid\bar{\boldsymbol{s}}_i\right)\pi_{i+1}\left(x_{i+1}\mid\boldsymbol{s}_{i+1}\right) \tag{279}$$

where every decision rule $\pi_i(X_i \mid \boldsymbol{S}_i)$, $i = 1, \ldots, H$, is a proper conditional distribution mapping from the domains of states $\boldsymbol{S}_i$ to action X_i. Probabilities of transition distributions $P_{\bar{\boldsymbol{x}}_i}(S_{i+1} \mid \bar{\boldsymbol{s}}_i)$ for $i = 0, \ldots, H-1$ are contained in the convex set defined in Eq. 271; values of the conditional reward $\mathbb{E}_{\bar{\boldsymbol{x}}_H}[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{s}}_H]$ are contained in the convex polytope $\boldsymbol{\mathcal{R}}$ defined in Eq. 272.

Solving for a policy $\pi^{(t)}$ and an optimistic SCM $\mathcal{M}^{(t)}$ in Eq. 269 is equivalent to solving a polynomial program with objective function $\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})]$ defined in Eq. 279 with transitional probabilities $P_{\bar{\boldsymbol{x}}_i}(S_{i+1} \mid \bar{\boldsymbol{s}}_i) \in \boldsymbol{\mathcal{P}}_i$, $i = 0, \ldots, H-1$, and reward mean $E_{\bar{\boldsymbol{x}}_K}[\mathcal{R}(\boldsymbol{Y})|\bar{\boldsymbol{s}}_K] \in \boldsymbol{\mathcal{R}}$; and probabilistic constraints $\sum_{x_i} \pi_i(x_i \mid \boldsymbol{s}_i) = 1$ and $\pi_i(x_i \mid \boldsymbol{s}_i) \geq 0$, for every $i = 1, \ldots, H$. There exists an efficient dynamic programming procedure to solve this polynomial program when the policy space Π satisfies the *perfect recall* condition (Koller and Friedman, 2009, Def. 23.5). In words, this conditions states that $\boldsymbol{S}_i \cup \{X_i\} \subseteq \boldsymbol{S}_j$ whenever $i < j$, which means that the agent does not forget the previous decision or information it once had.[44] An optimistic policy $\pi^{(t)}$ is obtainable by solving following extended Bellman equations, for $\forall i = 1, \ldots, H$,

$$Q^*(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_{i-1}) = \max_{x_i} \left\{ \max_{P_{\bar{\boldsymbol{x}}_i}(\cdot|\bar{\boldsymbol{s}}_i) \in \boldsymbol{\mathcal{P}}_i} \left\{ \sum_{s_{i+1}} Q^*(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i) P_{\bar{\boldsymbol{x}}_i}(s_{i+1}|\bar{\boldsymbol{s}}_i) \right\} \right\},$$

$$\text{and} \quad Q^*(\bar{\boldsymbol{s}}_H, \bar{\boldsymbol{x}}_{H-1}) = \max_{x_H} \max_{\mathbb{E}_{\bar{\boldsymbol{x}}_H}[\mathcal{R}(\boldsymbol{Y})|\bar{\boldsymbol{s}}_H] \in \boldsymbol{\mathcal{R}}} \mathbb{E}_{\bar{\boldsymbol{x}}_H}[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{s}}_H], \tag{280}$$

The inner maximum in the above equation is a linear program (LP) over the convex polytope $\boldsymbol{\mathcal{P}}_k$ (or $\boldsymbol{\mathcal{R}}$), which is solvable using by an iterative algorithm introduced by (Strehl and Littman, 2008). For grounding purposes, we provide the complete algorithm in Appendix. D.

Theorem 10 *Let $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ be a CDM where $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^H$ and $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto [0, 1]$. For any $\epsilon \geq 0$, the regret of UCB in SCM $\mathcal{M}^*$ after $T > 1$ episodes is bounded by*

$$R(T, \mathcal{M}^*) \leq \max_{\pi \in \Pi^{\text{o}}: \Delta_\pi > \epsilon} \frac{17^2 H^2 \left|\mathscr{D}(\boldsymbol{X} \cup \boldsymbol{S})\right| \log(T)}{\Delta_\pi} + \max_{\pi \in \Pi^{\text{o}}: \Delta_\pi \leq \epsilon} \Delta_\pi T + \frac{\pi^2}{6}. \tag{281}$$

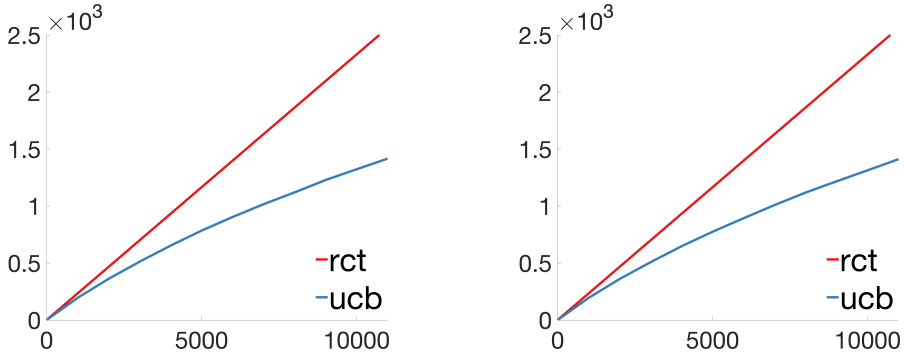
where $\Pi^{\text{o}} = \{\pi \in \Pi : \pi \text{ is deterministic}\}$ is the set of all deterministic policies in the policy space Π; and $\Delta_\pi = \mathbb{E}_{\pi^*}[Y; \mathcal{M}^*] - \mathbb{E}_\pi[Y; \mathcal{M}^*]$ is the gap from the optimal reward for any policy $\pi \in \Pi$. Moreover, fix $\epsilon = 0$. The regret of UCB could be further written as

$$R(T, \mathcal{M}^*) \leq \max_{\pi \in \Pi^{\text{o}}: \Delta_\pi > 0} \frac{17^2 H^2 \left|\mathscr{D}(\boldsymbol{X} \cup \boldsymbol{S})\right| \log(T)}{\Delta_\pi} + \frac{\pi^2}{6}. \tag{282}$$

■

Thm. 10 implies that Alg. 6 is able to achieve a sublinear regret $\mathcal{O}\left(H^2 \left|\mathscr{D}(\boldsymbol{X} \cup \boldsymbol{S})\right| \log(T)/\Delta\right)$ where H is the total number of actions (i.e., the decision horizon), $|\mathscr{D}(\boldsymbol{X} \cup \boldsymbol{S})|$ is the cardinality of the state-action domain; T is the total episodes of online interventions; and Δ is the gap in the expected reward between the second-best deterministic policy π and the optimal policy π evaluated in the underlying SCM $\mathcal{M}^*$. This means that UCB is able to converge and eventually obtain an optimal policy π^* as the total episodes of intervention T increases. Moreover, suppose $\mathcal{M}^*$ is an MAB model with K candidate arms, i.e., $H = 1$ and $|\mathscr{D}(\boldsymbol{X} \cup \boldsymbol{S})| = K$. The regret bound of Eq. 282 is equal to $\mathcal{O}(K \log(T)/\Delta)$, which matches the analytical result in Thm. 5.

44. In many real-world healthcare applications where the decision horizon H is low, the perfect recall assumption is quite natural and automatically satisfied. On the other hand, in some practical settings where the horizon H is high or even infinite, the policy space Π satisfying the perfect recall is high-dimensional. Planning in Π is computationally challenging even when parameters of the underlying SCM are fully known (Papadimitriou and Tsitsiklis, 1987).



(a) $\alpha_1 = 3, \alpha_2 = -3$ (b) $\alpha_1 = -12, \alpha_2 = -3$

Figure 26: Simulation results comparing UCB learner optimizing a 2-stage DTR model and RCT determining values of actions X_1, X_2 uniformly at random.

Experiment 5 *Fig. 26 shows the cumulative regret of* UCB$^+$ *in the 2-stage DTR environment* $\mathcal{M}^*$ *described in Example 12 with the coefficients* (α_1, α_2) *set to* $(3, -3)$ *and* $(-12, -3)$ *respectively. It takes as input the causal bounds over* $P_{X_1}(S_2 \mid S_1)$ *and* $\mathbb{E}_{X_1,X_2}[Y \mid S_1, S_2]$*, computed from the observational distribution* $P(S_1, X_1, S_2, X_2, Y)$*. As a baseline, we include a randomized controlled trials (*RCT*) deciding treatments* X_1, X_2 *uniformly at random, as introduced by (Murphy, 2005a), which extended the one-shot, classical treatment by (Fisher, 1935).*

One can see by inspection UCB *is able to achieve a sublinear regret in both DTR models. Simulations corroborate the analytical results* UCB *is able to eventually converge to an optimal policy* π^* *as the total number of trials* T *increases. As expected, the randomized strategy* RCT *performs the worst among all strategies due to the linear regret during exploration.* ■

5.3 Learning from Observational Data

Despite its performance guarantee, the online learning algorithm introduced in Alg. 6 does not make use of any knowledge in the observational distribution $P(\boldsymbol{V})$. When the NUC condition (Def. 13) holds in the underlying CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$, the state and actions' history $\bar{\boldsymbol{X}}_{i-1}, \bar{\boldsymbol{S}}_i$ blocks all backdoor paths every action X_i to any other variable in the causal diagram $\mathcal{G}$. One could thus estimate the transition distribution $P_{\bar{\boldsymbol{x}}_i}(\boldsymbol{S}_{i+1} \mid \bar{\boldsymbol{s}}_i)$, $i = 0, \ldots, H-1$ and reward $\mathbb{E}_{\bar{\boldsymbol{x}}_H}[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{s}}_H]$ using the corresponding conditional distribution $P(\boldsymbol{S}_{i+1} \mid \bar{\boldsymbol{x}}_i, \bar{\boldsymbol{s}}_i)$ and $\mathbb{E}[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H]$. The validity of the estimation procedure follows from Rule 2 of do-calculus (Def. 7). These estimations could then be directly transferred to "warm-start" UCB algorithm. However, issues of non-identifiability could arise in general settings where the NUC does not hold and no additional causal assumptions are provided.[45]

Given such challenges, we then consider the *partial identification* of the transition distributions and the expected reward from the observational distribution. Our first result bounds interventional transition probabilities $P_{\bar{\boldsymbol{x}}_i}(\boldsymbol{s}_{i+1} \mid \bar{\boldsymbol{s}}_i)$ from the observational distribution $P(\boldsymbol{V})$.

45. The non-identifiability of transition distributions $P_{\bar{\boldsymbol{x}}_i}(\boldsymbol{S}_{i+1} \mid \bar{\boldsymbol{s}}_i)$, $i = 2, \ldots, H-1$ and reward $\mathbb{E}_{\bar{\boldsymbol{x}}_H}[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{s}}_H]$ has been acknowledged in (Lee et al., 2019; Correa and Bareinboim, 2019).

Theorem 11 *Let $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ be a CDM where the policy space $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}$. For every $i = 1, \ldots, H-1$, $P_{\bar{\boldsymbol{x}}_i}(\boldsymbol{s}_{i+1} \mid \bar{\boldsymbol{s}}_i) \in [l_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}), r_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1})]$ where*

$$l_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}) = \frac{P(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i)}{\Gamma(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_{i-1})}, \qquad r_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}) = \frac{\Gamma(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i)}{\Gamma(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_{i-1})}. \tag{283}$$

and $\Gamma(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i)$ is a function of the observational distribution $P(\boldsymbol{V})$ defined as:

$$\Gamma(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i) = \begin{cases} P(\boldsymbol{s}_1) & \text{if } i = 0 \\ P(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i) - P(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_i) + \Gamma(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_{i-1}) & \text{if } i = 1, \ldots, H-1 \end{cases} \tag{284}$$

■

The upper bound in Eq. 283 could be written as:

$$\frac{\Gamma(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i)}{\Gamma(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_{i-1})} = \frac{P(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i) - P(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_i) + \Gamma(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_{i-1})}{\Gamma(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_{i-1})} \tag{285}$$

$$= 1 - \frac{P(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_i) - P(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i)}{\Gamma(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_{i-1})} \tag{286}$$

Considering the denominator, note that the gap $P(\bar{\boldsymbol{s}}_i, \bar{\boldsymbol{x}}_i) - P(\bar{\boldsymbol{s}}_{i+1}, \bar{\boldsymbol{x}}_i) > 0$ whenever observational probabilities $P(\boldsymbol{s}, \boldsymbol{x}) > 0$ are positive for all realizations of the state-action pair. This means that the causal bounds in Thm. 11 are generally informative, i.e., strictly contained in the real interval $[0, 1]$. The bounds following Thm. 11 can be seen as a generalization of the natural ones given in Thm. 8 to the sequential settings with multiple actions $H > 1$. The following example illustrates this connection.

Example 44 *Consider the 2-stage DTR model $\mathcal{M}^*$ described in Example 12. We will bound the interventional distribution $P_{x_1}(S_2 \mid s_1)$ from the observational distribution $P(S_1, X_1, S_2, X_2, Y)$. Applying Thm. 11, we obtain the lower bound*

$$P_{x_1}(s_2 \mid s_1) \geq \frac{P(s_1, s_2, x_1)}{\Gamma(s_1)} \tag{287}$$

$$\geq P(s_2, x_1 \mid s_1) \tag{288}$$

The last step follows from $\Gamma(s_1) = P(s_1)$. Similarly, function $\Gamma(s_1, s_2, x_1)$ could be written as:

$$\Gamma(s_1, s_2, x_1) = P(s_1, s_2, x_1) - P(s_1, x_1) + \Gamma(s_1) \tag{289}$$

$$= P(s_1, s_2, x_1) - P(s_1, x_1) + P(s_1) \tag{290}$$

Therefore, we could obtain the upper bound

$$P_{x_1}(s_2 \mid s_1) \leq \frac{\Gamma(s_1, s_2, x_1)}{\Gamma(s_1)} \tag{291}$$

$$\leq \frac{P(s_1, s_2, x_1) - P(s_1, x_1) + P(s_1)}{P(s_1)} \tag{292}$$

$$\leq P(s_2, x_1 \mid s_1) - P(x_1 \mid s_1) + 1 \tag{293}$$

The above bounds could be seen as an application of natural bounds (Thm. 8) with action X_1 and outcome S_2 conditioning on the covariate S_1. We evaluate transition probabilities $P_{x_1}(s_2 \mid s_1)$ in the underlying SCM $\mathcal{M}^*$, compute their corresponding causal bounds in Table 13a. ■

S_1	X_1	S_2	P	l	r	S_1	X_1	S_2	P	l	r
0	0	0	0.4750	0.1021	0.8872	1	0	0	0.4502	0.0069	0.9915
0	0	1	0.5250	0.1128	0.8979	1	0	1	0.5498	0.0085	0.9931
0	1	0	0.4502	0.3534	0.5683	1	1	0	0.4256	0.4190	0.4344
0	1	1	0.5498	0.4317	0.6466	1	1	1	0.5744	0.5656	0.5810

(a) $P_{X_1}(S_2 \mid S_1)$

S_1	X_1	S_2	X_2	$\mathbb{E}$	l	r	S_1	X_1	S_2	X_2	$\mathbb{E}$	l	r
0	0	0	0	0.7851	0.0587	0.9779	1	0	0	0	0.2149	0.0007	0.9959
0	0	0	1	0.9846	0.0333	0.9990	1	0	0	1	0.7851	0.0014	0.9992
0	0	1	0	0.7851	0.0138	0.9963	1	0	1	0	0.2149	0.0002	0.9991
0	0	1	1	0.7851	0.0744	0.9663	1	0	1	1	0.2149	0.0009	0.9934
0	1	0	0	0.2149	0.1279	0.6254	1	1	0	0	0.0008	0.0007	0.2417
0	1	0	1	0.9846	0.1170	0.9975	1	1	0	1	0.2149	0.0288	0.8232
0	1	1	0	0.2149	0.0490	0.8917	1	1	1	0	0.0008	0.0002	0.7897
0	1	1	1	0.7851	0.4021	0.8917	1	1	1	1	0.0154	0.0102	0.2472

(b) $\mathbb{E}_{X_1,X_2}[Y \mid S_1, S_2]$

Table 13: Interventional distributions $P_{X_1}(S_2 \mid S_1)$ and $\mathbb{E}_{X_1,X_2}[Y \mid S_1, S_2]$ and their causal bounds defined by a 2-stage DTR environment described in Example 12.

Similarly, one could bound conditional rewards $\mathbb{E}_{\bar{\boldsymbol{x}}_H}[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{s}}_H]$ from the observational data.

Theorem 12 *Let $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ be a CDM where the policy space $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}$ and the reward function $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto [0, 1]$. $\mathbb{E}_{\bar{\boldsymbol{x}}_H}[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{s}}_H] \in [l_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H), r_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H)]$ where*

$$
\begin{aligned}
l_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H) &= \frac{E[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{s}}_H, \bar{\boldsymbol{x}}_H] P(\bar{\boldsymbol{s}}_H, \bar{\boldsymbol{x}}_H)}{\Gamma(\bar{\boldsymbol{s}}_H, \bar{\boldsymbol{x}}_{H-1})} \\
r_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H) &= 1 - \frac{(1 - E[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{s}}_H, \bar{\boldsymbol{x}}_H]) P(\bar{\boldsymbol{s}}_H, \bar{\boldsymbol{x}}_H)}{\Gamma(\bar{\boldsymbol{s}}_H, \bar{\boldsymbol{x}}_{H-1})}
\end{aligned}
\tag{294}
$$

Since the conditional reward $E[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{s}}_H, \bar{\boldsymbol{x}}_H] \in [0, 1]$, the bounds in Eq. 294 must be strictly contained in $[0, 1]$ whenever the observational probability $P(\bar{\boldsymbol{s}}_H, \bar{\boldsymbol{x}}_H)$ is positive, and are thus informative. The bounds developed so far are functions of the observational distribution $P(\boldsymbol{V})$, which is identifiable by the sampling process, and so generally can be estimated consistently. We could estimate causal bounds in Thms. 11 and 12 by the corresponding sample mean estimates. Standard concentration inequalities are applicable to control the uncertainties due to finite samples.

Algorithm 7 Upper Confidence Bound with Causal Bounds (UCB$^+$)

Require: a policy space $\Pi = \{\langle X_i, \boldsymbol{S}_i\rangle\}_{i=1}^{H}$, a reward function $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto [0, 1]$, causal bounds $[l_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}), r_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1})]$, $i = 1, \ldots, H-1$, and $[l_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H), r_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H)]$ for all state-action pairs $(\boldsymbol{x}, \boldsymbol{s}) \in \mathscr{D}(\boldsymbol{X} \cup \boldsymbol{S})$.

1: Let $\mathbb{M}_{\text{c}}$ be the set of all SCMs $\mathcal{M}$ with endogenous variables $\boldsymbol{V}$, and with the transition distribution $P_{\bar{\boldsymbol{x}}_i}(\boldsymbol{S}_{i+1} \mid \bar{\boldsymbol{s}}_i; \mathcal{M})$ compatible with bounds $[l_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}), r_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1})]$, $i = 0, \ldots, H-1$, and rewards $\mathbb{E}_{\bar{\boldsymbol{x}}_H}[Y \mid \bar{\boldsymbol{s}}_H; \mathcal{M}]$ compatible with $[l_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H), r_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H)]$, that is,

$$\forall i = 0 \ldots, H-1,\ l_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}) \leq P_{\bar{\boldsymbol{x}}_i}(\boldsymbol{s}_{i+1} \mid \bar{\boldsymbol{s}}_i; \mathcal{M}) \leq r_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}), \tag{299}$$

$$\text{and } l_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H) \leq \mathbb{E}_{\bar{\boldsymbol{x}}_n}[Y \mid \bar{\boldsymbol{s}}_n; \mathcal{M}] \leq r_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H). \tag{300}$$

2: **for all** episodes $t = 1, 2, \ldots$ **do**

3: Construct a set of plausible SCMs $\mathbb{M}_{t-1}(\delta)$ with $\delta = t^{-4}$ following Steps 2-4 of Alg. 6.

4: Find the optimal policy $\pi^{(t)}$ of an optimistic SCM $\mathcal{M}^{(t)} \in \mathbb{M}_{t-1}(\delta) \cap \mathbb{M}_{\text{c}}$ such that

$$\mathbb{E}_{\pi^{(t)}}\left[Y; \mathcal{M}^{(t)}\right] = \max_{\pi, \mathcal{M}} \mathbb{E}_{\pi}\left[Y; \mathcal{M}\right] \text{ s.t. } \pi \in \Pi, \mathcal{M} \in \mathbb{M}_{t-1}(\delta) \cap \mathbb{M}_{\text{c}}. \tag{301}$$

5: Perform do $\left(\pi^{(t)}\right)$ for episode t and receive observations $\boldsymbol{V}^{(t)}$.

6: **end for**

Example 45 *Continue with the 2-stage DTR model in Example 44. We also apply Thm. 12 to bound the expected reward* $\mathbb{E}_{x_1,x_2}[Y \mid s_1, s_2]$. *Precisely,*

$$\mathbb{E}_{x_1,x_2}[Y \mid s_1, s_2] \geq \frac{\mathbb{E}[Y \mid s_1, s_2, x_1, x_2]\, P(s_1, s_2, x_1, x_2)}{\Gamma(s_1, s_2, x_1)} \tag{295}$$

$$\geq \frac{\mathbb{E}[Y \mid s_1, s_2, x_1, x_2]\, P(s_1, s_2, x_1, x_2)}{P(s_1, s_2, x_1) - P(s_1, x_1) + P(s_1)} \tag{296}$$

The last step follows from the evaluation of $\Gamma(s_1, s_2, x_1)$ *in Eq. 290. Similarly,*

$$\mathbb{E}_{x_1,x_2}[Y \mid s_1, s_2] \leq 1 - \frac{(1 - \mathbb{E}[Y \mid s_1, s_2, x_1, x_2])\, P(s_1, s_2, x_1, x_2)}{\Gamma(s_1, s_2, x_1)} \tag{297}$$

$$\leq 1 - \frac{(1 - \mathbb{E}[Y \mid s_1, s_2, x_1, x_2])\, P(s_1, s_2, x_1, x_2)}{P(s_1, s_2, x_1) - P(s_1, x_1) + P(s_1)} \tag{298}$$

We evaluate the conditional reward $\mathbb{E}_{x_1,x_2}[Y \mid s_1, s_2]$ *directly in the underlying SCM* $\mathcal{M}^*$*, compute their corresponding causal bounds derived above and provide them in Table 13b.* ■

We are ready to introduce a generalized UCB$^+$ algorithm utilizing causal bounds in the sequential decision-making setting. Alg. 7 summarizes the details of its implementation. It takes as input arguments a policy space Π, and bounds over transition probabilities $[l_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}), r_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1})]$, $i = 1, \ldots, H-1$, and rewards $[l_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H), r_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H)]$ computed from the observational distribution $P(\boldsymbol{V})$, following the derivation in Thms. 11 and 12. More specifically, in Step 1, UCB$^+$ constructs a family $\mathbb{M}_{\text{c}}$ of plausible SCMs with transition distributions $P_{\bar{\boldsymbol{x}}_i}(\boldsymbol{s}_{i+1} \mid \bar{\boldsymbol{s}}_i) \in [l_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}), r_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1})]$ and with rewards $\mathbb{E}_{\bar{\boldsymbol{x}}_H}[Y \mid \bar{\boldsymbol{s}}_H] \in [l_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H), r_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H)]$ compatible with the provided causal bounds.

Since bounds in Thms. 11 and 12 are sound and the observational data are often abundant, this ensures that the underlying SCM $\mathcal{M}^* \in \mathbb{M}_c$ with high probability. For every episode t, it computes the optimal policy $\pi^{(t)}$ of an optimistic SCM $\mathcal{M}^{(t)}$ in set intersection $\mathbb{M}_t \cap \mathbb{M}_c$ (Step 4). The construction of the SCM family $\mathbb{M}_t$ follows Steps 2-4 of UCB defined in Alg. 6. Similar to the optimistic planning procedure described previously, $\pi^{(t)}$ could be obtained by solving a polynomial program with the objective Eq. 263, subject to interventional constraints of Eqs. 271 and 272, and additional constraints of Eqs. 299 and 300 imposed by causal bounds.

The causal bounds in over transition distributions $[l_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1}), r_{\bar{\boldsymbol{x}}_i}(\bar{\boldsymbol{s}}_{i+1})]$, $i = 1, \ldots, H-1$, and rewards $[l_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H), r_{\bar{\boldsymbol{x}}_H}(\bar{\boldsymbol{s}}_H)]$ also permits a partial identification strategy to bound the expected rewards of candidate policies $\pi \in \Pi$. Formally, the expected reward $\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})] \in [l_\pi, r_\pi]$ such that

$$l_\pi = \min_{\mathcal{M} \in \mathbb{M}_c} \mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}], \qquad r_\pi = \max_{\mathcal{M} \in \mathbb{M}_c} \mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}]. \tag{302}$$

where $\mathbb{M}_c$ is the set of all SCMs compatible with constraints imposed by causal bounds. Interestingly, these bounds $[l_\pi, r_\pi]$ also characterize conditions under which the confounded observational data accelerate the performance of online learning algorithms.

Theorem 13 *Let $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ be a CDM where $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^H$ and $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto [0, 1]$. The regret of UCB$^+$ in SCM $\mathcal{M}^*$ after $T > 1$ episodes is bounded by*

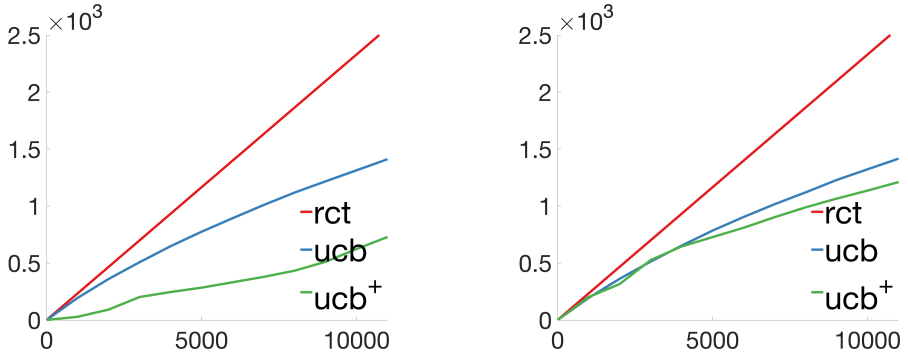
$$R(T, \mathcal{M}^*) \leq \max_{\pi \in \Pi^o : \substack{\Delta_\pi > 0, \\ r_\pi \geq \mu^*}} \frac{17^2 H^2 |\mathscr{D}(\boldsymbol{X} \cup \boldsymbol{S})| \log(T)}{\Delta_\pi} + \frac{\pi^2}{6}. \tag{303}$$

where $\mu^* = \mathbb{E}_{\pi^*}[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*]$ is the expected reward of an optimal policy $\pi^* \in \Pi$. ■

Thm. 13 implies that UCB$^+$, utilizing the observational data, consistently dominates UCB (Alg. 6) in terms of the performance. Broadly, it enjoys the same asymptotic regret bound as UCB, provided in Thm. 10. When the causal bounds are informative, i.e., there exist some suboptimal policies π with the upper bound $r_\pi < \mu_{\pi^*}$ smaller than the expected reward of the optimal policy π^*, UCB$^+$ is able to outperform UCB that learns from scratch, without using any prior observations. For instance, consider a multi-armed bandit model with action $|\mathscr{D}(X)| = K$ and states $\boldsymbol{S} = \emptyset$. The regret bound of UCB$^+$ is $\mathcal{O}(K \log(T)/\Delta_x)$ where Δ_x is the smallest gap among sub-optimal arms x with causal upper bound $r_x \geq \mu^*$. The improvement condition matches the analytical result in Thm. 9, derived for the special case of MAB models.

Experiment 6 *Fig. 27 shows the cumulative regret of UCB$^+$ in the 2-stage DTR environment $\mathcal{M}^*$ described in Example 12. The setup is the same as in Experiment. 5. It takes as input the causal bounds over $P_{X_1}(S_2 \mid S_1)$ and $\mathbb{E}_{X_1, X_2}[Y \mid S_1, S_2]$, computed from the observational distribution $P(S_1, X_1, S_2, X_2, Y)$. Simulation results show the significant disparity between the performance of UCB$^+$ and UCB for $(\alpha_1, \alpha_2) = (3, -3)$. In this case, the causal bound for the expected reward of a suboptimal policy $\pi = (X_1 \leftarrow 1, X_2 \leftarrow 0)$ is $r = 0.6283$, which is smaller than the optimal expected reward $\mu_{\pi^*} = 0.6757$. On the other hand, when the coefficients $(\alpha_1, \alpha_2) = (-12, -3)$ and the causal bound $r = 0.9875 > \mu_{\pi^*}$, the performance of UCB$^+$ and UCB coincides. These results corroborate the theory that the learning strategy of UCB$^+$ enjoys no negative impact.* ■

Table 14 summarizes online learning and offline-to-online learning algorithms studied so far in Secs. 4.1 and 5. These algorithms could be categorized into two lines of learning strategies: UCB



(a) $\alpha_1 = 3, \alpha_2 = -3$ (b) $\alpha_1 = -12, \alpha_2 = -3$

Figure 27: Simulation results comparing UCB$^+$ learner augmented with causal bounds over the expected rewards, standard UCB, and RCT determining values of action uniformly at random.

Decision Horizon	Algorithm	Regret Bound
$H = 1$	UCB (Alg. 3)	$\mathcal{O}\left(\lvert\mathscr{D}(X)\rvert \log(T)/\Delta\right)$
	UCB$^+$ (Alg. 5)	$\mathcal{O}\left(\lvert\mathscr{D}(X)\rvert \log(T)/\Delta_*\right)$
$H \geq 2$	UCB (Alg. 6)	$\mathcal{O}\left(H^2 \lvert\mathscr{D}(\boldsymbol{X} \cup \boldsymbol{S})\rvert \log(T)/\Delta\right)$
	UCB$^+$ (Alg. 7)	$\mathcal{O}\left(H^2 \lvert\mathscr{D}(\boldsymbol{X} \cup \boldsymbol{S})\rvert \log(T)/\Delta_*\right)$

Table 14: Summary of UCB and UCB$^+$ studied in Secs. 4.1 and 5. The performance gap $\Delta \leq \Delta_*$.

is an online algorithm that does not utilize any observational data, and UCB$^+$ is an offline-to-online algorithm that leverages on observational data through causal bounds. More specifically,

- Consider first when the decision horizon $H = 1$ and the input covariates $\boldsymbol{S} = \emptyset$. In this case, UCB achieves a regret bound $\mathcal{O}\left(\lvert\mathscr{D}(X)\rvert \log(T)/\Delta\right)$ where Δ is the smallest performance gap between the optimal arm x^* and a suboptimal arm x. Alg. 3 shows its implementation. On the other hand, UCB$^+$ utilizes the causal bounds and achieves a regret bound $\mathcal{O}\left(\lvert\mathscr{D}(X)\rvert \log(T)/\Delta_*\right)$ where Δ_* the smallest performance gap between the optimal arm x^* and a suboptimal arm x with a causal bound $r_x \geq \mu_{x^*}$. Alg. 5 shows its implementation. Since by definition, the performance gap $\Delta \leq \Delta_*$,UCB$^+$ performs at least as well as UCB.

- We also studied settings where the decision horizon $H \geq 2$ and every actions X_i is associated with a set of input covariates $\boldsymbol{S}_i$ for $i = 1, \ldots, H$. UCB, described in Alg. 6, is able to achieve a regret bound $\mathcal{O}\left(H^2 \lvert\mathscr{D}(\boldsymbol{X} \cup \boldsymbol{S})\rvert \log(T)/\Delta\right)$ where Δ is the smallest performance gap between the optimal policy π^* and a suboptimal deterministic policy π. Meanwhile, UCB$^+$ (Alg. 7) exploits the causal bounds and enjoys a regret bound $\mathcal{O}\left(H^2 \lvert\mathscr{D}(\boldsymbol{X} \cup \boldsymbol{S})\rvert \log(T)/\Delta_*\right)$ where Δ_* is the smallest performance gap between the optimal policy x^* and a suboptimal deterministic policy x with a causal bound $r_\pi \geq \mu_{\pi^*}$. Again, since the performance gap $\Delta \leq \Delta_*$, UCB$^+$ generally outperforms UCB in sequential settings.

After all, our analytical results reveal that causal bounds are robust to the confounding bias in the observational data, and could consistently improve the performance of online learners.

6. Mixed Policy Learning: Where to Intervene (CRL Task 2)

Agents are deployed in complex and uncertain environments, where they are exposed and need to process high volumes of information while being expected to operate efficiently, surgically, and safely. This requires the agent to identify an optimal policy to bring about a desirable state of affairs. A prevalent assumption in the literature, including the discussion in the previous sections, is that the action space is fixed. For example, it could be defined over variables $\boldsymbol{X}$, where $|\boldsymbol{X}| = k$. This implies that the agent will explore policies within the domain of $\boldsymbol{X}$, for example, encompassing 2^k possible configurations in the binary case.

In this section, we will relax the assumption that the policy scope is fixed and explore more flexible action spaces, including situations where the agent is not required to perform interventions. This is motivated by the observation that such a strategy, while viable in controlled, artificial environments (e.g., actions executed in a simulator or gaming scenarios), where interventions are harmless, becomes less ideal and sometimes infeasible in real-world settings due to their potentially harmful side effects.

Another complementary property, perhaps surprisingly, in non-Markovian causal systems is that controlling all intervenable variables, denoted as $do(\boldsymbol{X} \leftarrow \boldsymbol{x})$, does not necessarily lead to an optimal policy. In particular, we will show that in certain settings, a partial intervention, where $do(\boldsymbol{X}' \leftarrow \boldsymbol{x}')$ with $\boldsymbol{X}' \subset \boldsymbol{X}$, can outperform the case where full control is exerted. In such systems, a larger search would be required to examine all possible subsets of $\boldsymbol{X}$, including 3^k possible configurations, which will need to be evaluated in a systematic manner.

In real-world causal systems, it's interesting to recognize that full controllability is not always necessary. For instance, natural mechanisms often govern the action variables $\boldsymbol{X}$ in many scenarios, meaning that forcing the variable to take some value by external interventions might lead to undesirable effects. Robots inevitably obey the laws of physics, such as inertia and gravity, which makes their joints move naturally when their gears are disengaged; similarly, physicians treat patients based on their experience while adhering to varying rules and regulations by location. Therefore, in certain situations, it may be sufficient to intervene in only a subset of variables among $\boldsymbol{X}$, allowing the remaining variables to vary according to their natural dynamics, as determined by the underlying mechanism f_X. Fig. 28 illustrates these dynamics, where the agent performs interventions over different sets of variables in each episode, and sometimes just acts naturally.

Towards formalizing this setting, we put these observations together and define a *mixed policy space*, which is a collection of policy spaces in which each action and context are defined as subsets of intervenable variables $\boldsymbol{X}^\star$ and context variables $\boldsymbol{S}^\star$, respectively. That is,

$$\Pi_{\text{MIX}} = \{\{\langle X, \boldsymbol{S}_X\rangle\}_{X\in\boldsymbol{X}'} : \boldsymbol{X}' \subseteq \boldsymbol{X}^\star, \boldsymbol{S}_X \subseteq \boldsymbol{S}^\star\}.$$

Every policy space in Π_{MIX} lies between two extremes — observational and experimental policies.

Definition 16 (Mixed Policy Space & Policy) *Let $\mathcal{G}$ be a causal diagram, $Y \in \boldsymbol{V}(\mathcal{G})$ be a reward variable. $\boldsymbol{X}^\star \subseteq \boldsymbol{V} \setminus \{Y\}$ be a set of intervenable variables, and $\boldsymbol{S}^\star \subseteq \boldsymbol{V} \setminus \{Y\}$ be a set of context variables. A mixed policy space Π_{MIX} is the collection of policy spaces where each policy space $\Pi \in \Pi_{\text{MIX}}$ is defined with actions $\boldsymbol{X} \subseteq \boldsymbol{X}^\star$ and contexts $\boldsymbol{S} \subseteq \boldsymbol{S}^\star$ such that $\Pi = \{\langle X_i, \boldsymbol{S}_i\rangle\}_{i|X_i\in\boldsymbol{X}}$, $\boldsymbol{S} =$*

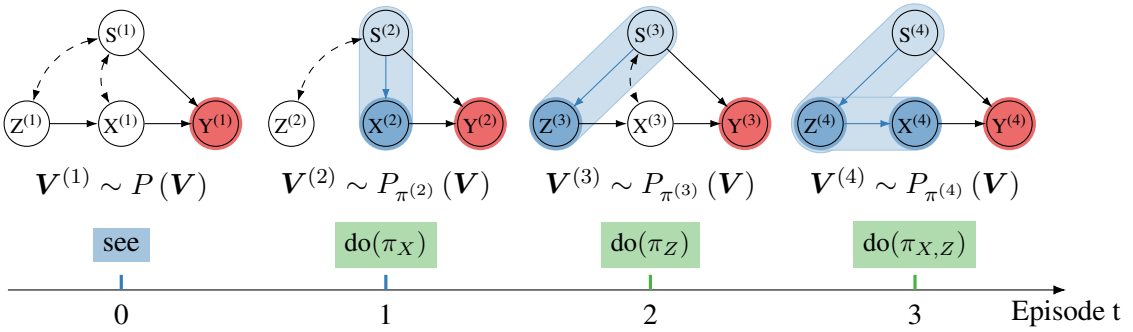


Figure 28: Temporal diagram showing the dynamics of a mixed policy learning while the agent interacts with the environment with different policy scopes at each episode.

$\bigcup_{i|X_i \in \boldsymbol{X}} \boldsymbol{S}_i$, *and* $\mathcal{G}_\Pi$ *is a DAG. Given a mixed policy space* Π_{MIX} *with respect to* $\langle \mathcal{G}, Y, \boldsymbol{X}^\star, \boldsymbol{S}^\star \rangle$, *a* mixed policy $\pi \in \Pi \in \Pi_{\text{MIX}}$ *is a policy* π *following the policy space* Π. ■

For simplicity, we may use $\pi \in \Pi_{\text{MIX}}$, which is the shorthand notation for $\pi \in \Pi \in \Pi_{\text{MIX}}$. The following task signature characterizes this learning setting involved in a mixed policy space:

$$\mathcal{T}_{\text{MIX}} = \left\langle \mathcal{I} = \text{do}, \mathcal{A} = \mathcal{G}, \Pi_{\text{MIX}} = \left\{ \{ \langle X_i, \boldsymbol{S}_i \rangle \mid \boldsymbol{S}_i \subseteq \boldsymbol{S}^\star \}_{X_i \in \boldsymbol{X}} \right\}_{\boldsymbol{X} \subseteq \boldsymbol{X}^\star}, \mathcal{R} = \mathscr{D}(Y) \mapsto \mathbb{R} \right\rangle.$$

Note that the policy space is not fixed but an element of a mixed policy space. This means that the agent will search for a policy π^* such that

$$\pi^* = \underset{\pi \in \Pi_{\text{MIX}}}{\arg\max}\, \mathbb{E}_\pi^{\mathcal{M}^*} \left[\mathcal{R}(\boldsymbol{Y}) \;\middle|\; \mathcal{G},\ \mathcal{D}_{\text{exp}} \sim P_{\boldsymbol{x}}(\boldsymbol{V}) \right], \tag{304}$$

where the distinct feature of the task is the mixed policy scope.

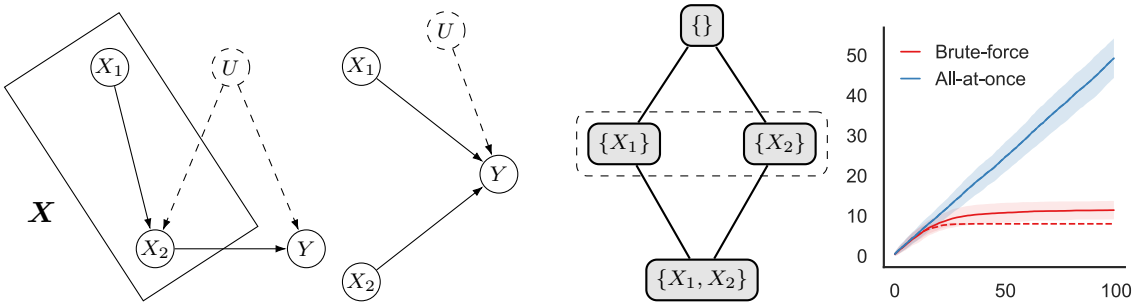
In Section 6.1, we introduce the marginal case of mixed policy spaces, where no context variables are present. Even without considering contexts, the problem of deciding where the agent should intervene in the system is challenging and sets the ground for further explorations. We then investigate in Section 6.2 mixed policy spaces with context variables, or where the agent should intervene (do) and look (see) to determine the optimal policy.

6.1 Mixed Policy with No Context

In this section, we investigate how an agent should behave to efficiently identify an optimal action given a mixed policy space and an underlying causal diagram. For simplicity, we focus on MAB settings where each policy $\pi \in \Pi_{\text{mix}}$ is an experimental policy $(\pi_1, \ldots, \pi_{|\boldsymbol{X}'|})$ over a subset of actions $\boldsymbol{X}' \subseteq \boldsymbol{X}^\star$ with each decision rule $\pi(x)$ involved in an empty context. Thus, $\pi(\boldsymbol{x}') \in \Pi_{\text{mix}}$ for $\boldsymbol{x}' \in \mathscr{D}(\boldsymbol{X}')$ and $\boldsymbol{X}' \subseteq \boldsymbol{X}^\star$. The following example illustrates the challenge of this task.

Example 46 (Where to intervene) *Consider a MAB environment described by an SCM*

$$\mathcal{M}^* = \langle \boldsymbol{U} = \{U_1, U_2\}, \boldsymbol{V} = \{X_1, X_2, Y\}, \mathscr{F}, P(U_1, U_2) \rangle, \tag{305}$$



(a) (b) (c) (d)

Figure 29: (a) True causal diagram $\mathcal{G}$ for environment $\mathcal{M}^*$. (b) Hypothesized model in the agent's mind, after intervention. (c) Structure of the action space. (d) Two agents' cumulative regret with Thompson sampling (solid line) and UCB (dashed line) together with shaded areas representing 95% confidence interval. Two lines for the All-at-once agent are overlapped.

where the endogenous variables $\boldsymbol{V}$ are all binary. The causal mechanisms are the following:

$$\mathscr{F} = \begin{cases} X_1 & \leftarrow U_1, \\ X_2 & \leftarrow X_1 \oplus U_2, \\ Y & \leftarrow X_2 \oplus U_2, \end{cases} \tag{306}$$

and the exogenous distribution:

$$P(U_1) = P(U_2) = 1/2. \tag{307}$$

The causal diagram associated with $\mathcal{M}^*$ is shown in Fig. 29a.

We now consider an agent deployed in $\mathcal{M}^*$ with the goal of optimizing the outcome variable Y while being capable of intervening on (or controlling) variables X_1, X_2. Specifically, the goal of the agent is to find

$$\pi^* = \arg\max_{x_1, x_2} \mathbb{E}_{X_1 \leftarrow x_1, X_2 \leftarrow x_2}[Y; \mathcal{M}^*] \tag{308}$$

The structure of the mixed policy space is shown in Fig. 29c, which highlights the various policy scopes available. In particular, this space represents the power set of action variables $\boldsymbol{X}^\star = \{X_1, X_2\}$, which entails $2^{|\boldsymbol{X}^\star|}$ possible intervention sets. This setting may be translated to a traditional MAB instance such that each arm corresponds to intervening on a subset of $\{X_1, X_2\}$ to a specific value, which results in 9 arms in this case.[46] [47]

We start our analysis with an agent that is oblivious to the causal structure underlying the action space and adheres to a strict experimentalist approach. In other words, this means that it will abstract away the causal diagram and perform interventions on one action variable $\boldsymbol{X} = \{X_1, X_2\}$.

46. $\{\text{do}(\emptyset), \text{do}(X_1 \leftarrow 0), \text{do}(X_1 \leftarrow 1), \text{do}(X_2 \leftarrow 0), \text{do}(X_2 \leftarrow 1), \text{do}(X_1 \leftarrow 0, X_2 \leftarrow 0), \text{do}(X_1 \leftarrow 0, X_2 \leftarrow 1), \text{do}(X_1 \leftarrow 1, X_2 \leftarrow 0), \text{do}(X_1 \leftarrow 1, X_2 \leftarrow 1)\}$

47. The same observation applies for MDPs or any more complex model, as discussed later on in the section.

We call the agent following such a strategy "all-at-once," since all variables are intervened together.[48] *Each intervention on $do(\boldsymbol{X})$ corresponds to interventions on the lower-level variables $\{do(X_1 \leftarrow 0, X_2 \leftarrow 0), do(X_1 \leftarrow 0, X_2 \leftarrow 1), do(X_1 \leftarrow 1, X_2 \leftarrow 0), do(X_1 \leftarrow 1, X_2 \leftarrow 1)\}$. Under an interventional regime, the causal structure compatible with this strategy is shown in Fig. 29b, which is clearly different from the environment's causal diagram.*

Despite what is in the agent's mind, or optimization function, it will still be evaluated by the underlying SCM $\mathcal{M}^*$. The natural question that arises here is whether it is okay to be oblivious to the pair $(\mathcal{G}, \mathcal{M})$; would it be sufficient to perform more interventions to make up for the ignorance of the causal structure? In other words, more samples from the $do(X_1 \leftarrow x_1, X_2 \leftarrow x_2)$ distribution should be sufficient to eventually learn an optimal policy?

This agent is deployed in the environment $\mathcal{M}^*$ and after some interventions, it is able to learn a policy and obtain the following reward:

$$\mathbb{E}_{x_1,x_2}[Y] = \mathbb{E}[x_2 \oplus U_{X_2,Y}] = 0.5x_2 + 0.5(1 - x_2) = 0.5. \tag{309}$$

This policy is, in fact, independent of the specific values of X_1 and X_2, and Y reaches its highest value at most 0.5 of the time. At this point, the expectation by some is that there is an issue of sample complexity, but not of asymptotic convergence. In other words, the agent can be oblivious to the causal structure, $\mathcal{G}$, compensating by accumulating more samples, but it would eventually be able to learn the optimal policy.

Now we examine an alternative policy within the mixed policy space in which the agent controls only one variable X_1. The evaluation of such a policy goes as follows:

$$\mathbb{E}_{x_1}[Y] = \mathbb{E}[(x_1 \oplus U_{X_2,Y}) \oplus U_{X_2,Y}] = x_1 \tag{310}$$

This means that if the goal is to keep Y as high as possible, the agent should perform an intervention $do(X_1 \leftarrow 1)$, which would imply that $Y = 1$ in each subsequent round.

The implication of such a result is that the strategy "all-at-once", oblivious to the environment's structure, will never converge, no matter how many interactions are allowed to the agent. We compare empirically this with an alternative strategy called "brute-force", which searches over the entire policy space, including all possible subsets of $\{X_1, X_2\}$. The performance of both agents is shown in Figure 29d, where the y-axis represents a cumulative regret. In fact, the all-at-once agent does not converge while the brute-force approach is able to find the optimal policy, since $do(X_1 \leftarrow 1)$ is inside the mixed policy space.

The question here is whether we can do better by leveraging the underlying causal invariances of $\mathcal{M}^*$, as represented in $\mathcal{G}$. We answer this question by examining the expected rewards over the entire mixed policy space. First, using Rule 3 of do-calculus, we note that $P(y \mid do(x_1, x_2)) = P(y \mid do(x_2))$, since X_1 has no effect on Y under intervention on X_2. That is, the corresponding expected rewards are equivalent, $\mu_{x_1,x_2} = \mu_{x_2}$, for any x_1 and x_2. Hence, $\mu^*_{X_1,X_2} = \mu^*_{X_2}$. Since the all-at-once strategy can be discarded, we examine 5 arms based on 3 intervention sets as follows:

$$\mathbb{E}[Y] = \mathbb{E}[(X_1 \oplus U_{X_2,Y}) \oplus U_{X_2,Y}] = \mathbb{E}[X_1] = 0.5 \tag{311}$$

$$\mathbb{E}_{x_1}[Y] = \mathbb{E}[(x_1 \oplus U_{X_2,Y}) \oplus U_{X_2,Y}] = x_1 \tag{312}$$

$$\mathbb{E}_{x_2}[Y] = \mathbb{E}[x_2 \oplus U_{X_2,Y}] = 0.5x_2 + 0.5(1 - x_2) = 0.5. \tag{313}$$

48. Formally, this can be thought of as a specific instance of a cluster causal diagram, an object that has been studied in the literature; for a more detailed discussion, refer to (Anand et al., 2021).

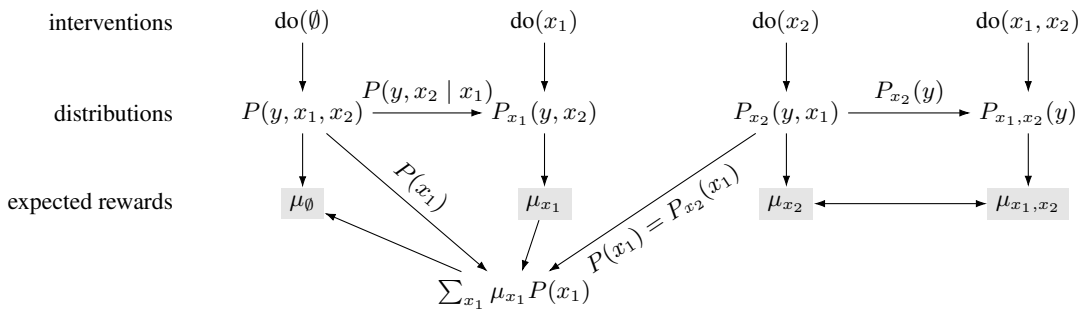


Figure 30: Relationships among quantities such as probability distributions and expected rewards arising in the mixed policy relative to causal model in Fig. 29a.

Therefore, the optimal action is do$(X_1 \leftarrow 1)$ *for the model with* $\mu^*_{X_1 \leftarrow 1} = 1$. *Again, intervening* $\{X_1, X_2\}$ *will incur regrets and cannot converge to the optimal solution.*

To explain this further, we now investigate the relationships among interventional probabilities in the causal model in Fig. 29a. The observational probability $P(y) = P(y \mid do(\emptyset))$ *can be viewed as a convex combination of* $\{P(y \mid do(x_1))\}_{x_1 \in \mathscr{D}(X_1)}$,

$$P(y) = \sum_{x_1} P(y \mid x_1) P(x_1) = \sum_{x_1} P_{x_1}(y)\, P(x_1). \tag{314}$$

That is, $\mu_\emptyset = \sum_{x_1} \mu_{x_1} P(x_1)$. *By replacing* μ_{x_1} *to* $\mu^*_{X_1} = \max_{x_1} \mu_{x_1}$, *then,*

$$\mu_\emptyset = \sum_{x_1} \mu_{x_1} P(x_1) \leq \sum_{x_1} \mu^*_{X_1} P(x_1) = \mu^*_{X_1}. \tag{315}$$

This equation holds for any model conforming to the causal diagram in Fig. 29a, namely, it affects whether the agent should play some arms since playing non-optimal arms will incur regrets. At this point, playing the arms over X_2 *is preferred to the arms over both* X_1 *and* X_2 *(i.e.,* $do(X_1, X_2)$*), since it minimizes the number of arms that need to be played to find the optimal, and* $do(X_1)$ *is preferred to* $do()$ *as* $\mu_\emptyset$ *cannot be strictly better than the best achievable expected reward obtainable by intervening on* X_1 *to* x_1^*.

Regarding the superiority of intervening on a set of variables over other set of variables, a natural question is, then, whether the comparisons among μ_{x_2} *and* μ_{x_1} *can be made as well, noting that* $\mu^*_{X_2} < \mu^*_{X_1}$ *in* $\mathcal{M}^*$. *In fact, we can show that the inequality* $\mu^*_{X_2} > \mu^*_{X_1}$ *is also realizable. To witness, consider an SCM* $\mathcal{M}'$ *identical to the one defined previously (Eq. 306) but for* Y*'s mechanism:*

$$f_Y \leftarrow X_2 + U_{X_2,Y}. \tag{316}$$

Then, we can evaluate the expected rewards in $\mathcal{M}'$ *as follows:*

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[(X_1 \oplus U_{X_2,Y}) + U_{X_2,Y}] = 1 \\
\mathbb{E}_{x_1}[Y] &= \mathbb{E}[(x_1 \oplus U_{X_2,Y}) + U_{X_2,Y}] = 0.5x_1 + 0.5(2 - x_1) = 1 \\
\mathbb{E}_{x_2}[Y] &= \mathbb{E}[x_2 + U_{X_2,Y}] = x_2 + 0.5
\end{aligned}$$

Hence, the optimal action is do$(X_2 \leftarrow 1)$ with $\mu^*_{X_2 \leftarrow 1} = 1.5$. This demonstrates the impossibility that arises in some cases of deciding a priori to prefer one interventional scope over the other solely based on the causal diagram, and this depends on the specific instantiation of the environment. ■

Considering the example, we note that ignoring the underlying causal structure, and the interplay between action space and reward, may result in a suboptimal performance due to playing such regret-incurring arms. If one is negligent to the influence of unobserved confounder and simply chooses to intervene on every variable, do(x_1, x_2), or to intervene on the one closest to Y, do(x_2), it is possible that the agent will never converge to the optimal arm, e.g., do(x_1^*).

Exploring this example, we have shown the existence of equivalence classes among actions with respect to their expected rewards. Also, certain partial-orders emerge among subsets of action variables with respect to their optimal expected rewards. Also, the expected reward of an action is related to other actions, e.g., an observational probability written with probabilities from do(x_1) and do(x_2), $P(y) = \sum_{x_1} P_{x_1}(y)P(x_1) = \sum_{x_1} P_{x_1}(y)P_{x_2}(x_1)$. Figure 30 illustrates relationships among different distributions and their rewards. Four different interventions are shown with their distributions where some distributions can yield other distributions, e.g., $P_{x_1}(y, x_2)$ from $P(y, x_1, x_2)$. Further, the expected reward for observation (mentioned above) can be represented as an expression made of probabilities from other arms. We will later see that such a formula improves the performance of online learners. We now more formally investigate these phenomena as studied in (Lee and Bareinboim, 2019a, 2018a, 2020).

6.1.1 STRUCTURAL PROPERTIES IN MIXED POLICY LEARNING IN A MAB SETTING

Here, we provide three structural properties emerging in a bandit setting with a mixed policy. These properties among different actions arise due to the shared causal mechanisms and can be understood through do-calculus and related machinery.

Property 1. Equivalence among actions Do-calculus provides rules to examine equivalence relationships in the space of conditional interventional distributions. Hence, it naturally partitions the space into equivalence classes. In particular, we focus on Rule 3, which ascertains a graphical condition such that a set of interventions does not have an effect on the outcome variable, i.e., $P(y \mid \text{do}(\boldsymbol{x}, \boldsymbol{z}), \boldsymbol{w}) = P(y \mid \text{do}(\boldsymbol{x}), \boldsymbol{w})$. Since actions correspond to interventions (including the null intervention) and there is no contextual information, we consider examining $P(y \mid \text{do}(\boldsymbol{x}, \boldsymbol{z})) = P(y \mid \text{do}(\boldsymbol{x}))$ through $(Y \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{X})$ in $\mathcal{G}_{\overline{\boldsymbol{X} \cup \boldsymbol{Z}}}$, which implies that $\mu_{\boldsymbol{x},\boldsymbol{z}} = \mu_{\boldsymbol{x}}$. If d-separation holds in the manipulated graph, this condition implies that it is sufficient to play only one action among actions in the equivalence class regarding finding an optimal arm efficiently. In an online learning setting where its objective is minimizing a cumulative regret, it is desired to play a smaller subset of arms given a set of arms as far as the subset contains the best arm. Against this background, we define a minimal intervention set.

Definition 17 (Minimal Intervention Set (MIS)) *A subset of action variables $\boldsymbol{X}' \subseteq \boldsymbol{X}^\star$ is said to be a minimal intervention set relative to $\mathcal{G}$, $\boldsymbol{X}^\star$, and Y if there is no proper subset $\boldsymbol{X}'' \subset \boldsymbol{X}'$ such that $\mu_{\boldsymbol{x}''} = \mu_{\boldsymbol{x}'}$ for every SCM conforming to $\mathcal{G}$ and $\boldsymbol{x}'' \in \mathscr{D}(\boldsymbol{X}'')$ consistent with $\boldsymbol{x}'$.* ■

Whether a subset of action variables $\boldsymbol{X}' \subseteq \boldsymbol{X}^\star$ is an MIS can be examined through a rather simple procedure involving in an ancestral relationship.

Proposition 5 (Minimality) *A set of variables $\boldsymbol{X}' \subseteq \boldsymbol{X}^\star$ is a minimal intervention set for $\mathcal{G}$ with respect to Y if and only if $\boldsymbol{X}' \subseteq an(Y)_{\mathcal{G}_{\overline{\boldsymbol{X}'}}}$.* ■

This characterization demonstrates that one can consider only directed edges among variables, not the unobserved variables, in acquiring MISes. Intervening nothing or a single variable (an ancestor of Y in $\mathcal{G}$) constitutes MISes. Further, intervening on $\boldsymbol{W} = pa(\boldsymbol{Z}) \setminus \boldsymbol{Z}$ is also an MIS for any $\boldsymbol{Z} \subseteq an(Y)_{\mathcal{G}}$ since each variable $W \in \boldsymbol{W}$ has a directed path towards Y without passing through the rest of $\boldsymbol{W}$, i.e., $\boldsymbol{W} \setminus \{W\}$.

Property 2. Partial-orders among minimal intervention sets We now explore the partial-orders among the subsets of $\boldsymbol{X}^\star$ within the MISes. Given a causal diagram $\mathcal{G}$, it is possible that intervening on some variables is *always* as good as intervening on another set of variables. Formally, there can be two different sets of variables $\boldsymbol{W}, \boldsymbol{Z} \subseteq \boldsymbol{X}^\star$ such that

$$\max_{\boldsymbol{w} \in \mathscr{D}(\boldsymbol{W})} \mu_{\boldsymbol{w}} \leq \max_{\boldsymbol{z} \in \mathscr{D}(\boldsymbol{Z})} \mu_{\boldsymbol{z}}$$

in every possible SCM conforming to $\mathcal{G}$. If that is the case, it would be unnecessary (and possibly harmful in terms of the sample efficiency) to play actions over $\mathscr{D}(\boldsymbol{W})$. We define Possibly-Optimal MIS, which incorporates the partial-orderedness among MISes denoting the optimal value for $\boldsymbol{X}' \subseteq \boldsymbol{X}^\star$ given an SCM by $\boldsymbol{x}'^*$.

Definition 18 (Possibly-Optimal Minimal Intervention Set (POMIS)) *Given $\mathcal{G}$, $\boldsymbol{X}^\star$ and Y, let $\boldsymbol{X}' \subseteq \boldsymbol{X}^\star$ be a MIS. If there exists an SCM conforming to $\mathcal{G}$ such that $\mu_{\boldsymbol{x}'^*} > \forall_{\boldsymbol{Z} \in \mathbb{Z} \setminus \{\boldsymbol{X}'\}} \mu_{\boldsymbol{z}^*}$, where $\mathbb{Z}$ is the set of MISes with respect to $\mathcal{G}$, $\boldsymbol{X}^\star$ and Y, then $\boldsymbol{X}'$ is a possibly-optimal minimal intervention set with respect to $\mathcal{G}$, $\boldsymbol{X}^\star$ and Y.* ■

To determine whether intervening on a subset of $\boldsymbol{X}^\star$ is a POMIS or not, one may list all possible partial-orders among MISes in a brute-force manner, and select those that are not dominated by any other MISes. However it is unclear whether we can compare two arbitrary MISes under what conditions and, further, whether such conditions are complete. As a starting point, we provide a way to obtain a single partial-order among two MISes. Consider intervening on an MIS $\boldsymbol{W}$. By basic algebra, we can express the expected reward for do$(\boldsymbol{w})$ for some $\boldsymbol{Z} \subseteq \boldsymbol{X}^\star$:

$$\mathbb{E}_{\boldsymbol{w}}[Y] = \sum_{\boldsymbol{z}} \mathbb{E}_{\boldsymbol{w}}[Y \mid \boldsymbol{z}] P_{\boldsymbol{w}}(\boldsymbol{z}).$$

If it is possible to exchange the observation $\boldsymbol{z}$ to an intervention $\boldsymbol{z}$ using Rule 2 of do-calculus, then the expression becomes

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{w}}[Y] &= \sum_{\boldsymbol{z}} \mathbb{E}_{\boldsymbol{z},\boldsymbol{w}}[Y] P_{\boldsymbol{w}}(\boldsymbol{z}) \\
&\leq \sum_{\boldsymbol{z}} \mathbb{E}_{(\boldsymbol{z},\boldsymbol{w})^*}[Y] P_{\boldsymbol{w}}(\boldsymbol{z}) \\
&= \mathbb{E}_{(\boldsymbol{z},\boldsymbol{w})^*}[Y] \\
&= \mathbb{E}_{(\boldsymbol{z}',\boldsymbol{w}')^*}[Y].
\end{aligned}$$

where $\boldsymbol{Z}' \cup \boldsymbol{W}'$ is an MIS corresponding to the intervention set $\boldsymbol{Z} \cup \boldsymbol{W}$ and * over $(\boldsymbol{z}, \boldsymbol{w})$ indicates the values for $(\boldsymbol{z}, \boldsymbol{w})$ maximizing the expectation.

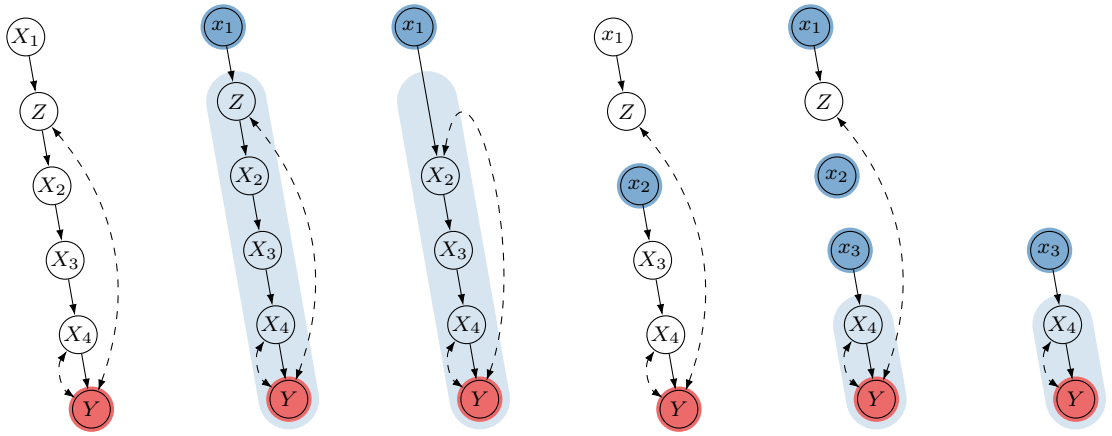


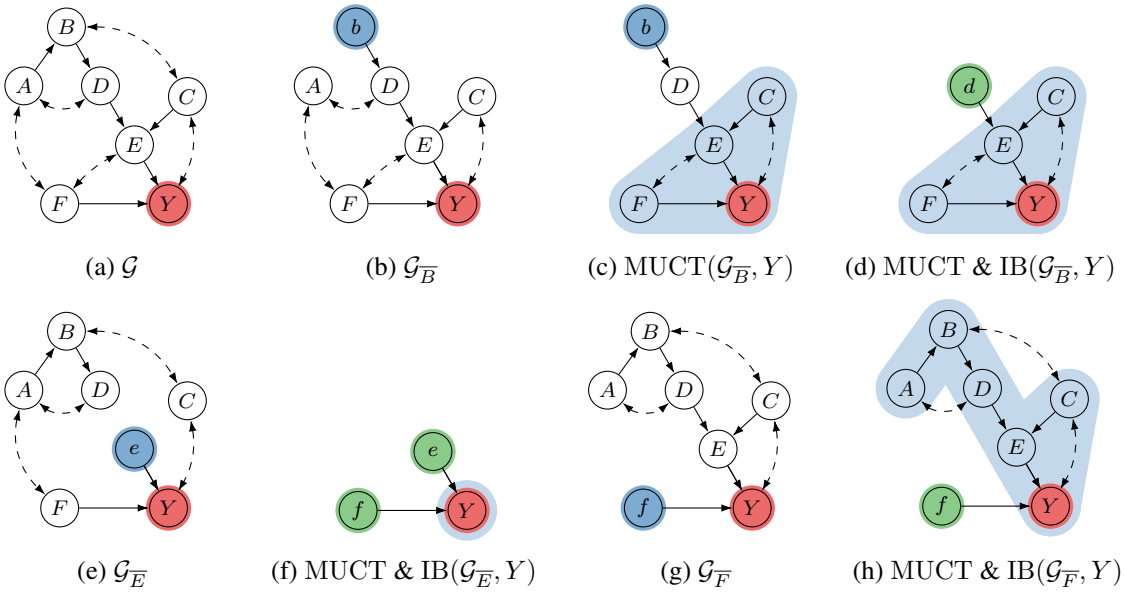
Figure 31: Illustrative examples demonstrating how partial-orders can be obtained from Figure 31a. (b and c) demonstrates no better intervention than $\text{do}(X_1)$ is obtained, (d to f) illustrates $\mu^*_{X_2} \leq \mu^*_{X_3}$. Light blue areas represent variables having a backdoor path to Y under the intervention.

To find such $\boldsymbol{Z} \subseteq \boldsymbol{X}^\star$ under the intervention on $\boldsymbol{W}$, we examine Rule 2 of do-calculus: $\boldsymbol{Z}$ must satisfy $(Y \perp\!\!\!\perp \boldsymbol{Z} \mid \boldsymbol{W})$ in $\mathcal{G}_{\overline{\boldsymbol{W}}\underline{\boldsymbol{Z}}}$, which is equivalent to $(Y \perp\!\!\!\perp \boldsymbol{Z})$ in $(\mathcal{G} \setminus \boldsymbol{W})_{\underline{\boldsymbol{Z}}}$. If no backdoor path from $\boldsymbol{Z}$ to Y exists in $\mathcal{G} \setminus \boldsymbol{W}$, then $\mu^*_{\boldsymbol{W}} \leq \mu^*_{\boldsymbol{W},\boldsymbol{Z}} = \mu^*_{\boldsymbol{W}',\boldsymbol{Z}'}$. One may iteratively apply this idea to find a set which leads to a higher expected reward when intervened on until stuck at a graph where every non-intervened subset of $\boldsymbol{X}^\star$ in the graph involves a backdoor path to Y.

A few examples are illustrated in Figure 31. In the example with a causal diagram where none is intervened on (Figure 31a), every X_i with $i > 1$ has a backdoor path through Z while X_4 is also directly confounded with Y. The light blue area in Figure 31b covers backdoor paths from a subset of $\boldsymbol{X}$ to Y in $\mathcal{G}_{\overline{X_1}}$. Thus, $\mu_\emptyset \leq \mu_{x_1}$ can be inferred. One may directly facilitate a graph obtained by projecting out variables neither $\boldsymbol{X}$ nor Y, and still backdoor paths can be equally examined in the resulting graph (Figure 31c). For intervening on X_2 (Figure 31d), both X_1 and X_3 have no backdoor path to Y (Figure 31e). Considering the minimality, we derive $\mu^*_{X_2} \leq \mu^*_{X_2,X_3,X_1} \leq \mu^*_{X_3}$ (Figure 31f). We can avoid considering X_1 in the beginning by excluding variables ineffective to Y under the intervention $\text{do}(x_2)$. We define a set of variables having backdoor paths to Y in a given graph:

Definition 19 (Minimal Unobserved-Confounders' Territory) *Given a diagram $\mathcal{G}$ and a node Y, let $\mathcal{H}$ be $\mathcal{G}[An(Y)_\mathcal{G}]$. A set of variables $\boldsymbol{T} \subseteq \boldsymbol{V}(\mathcal{H})$ containing Y is called a UC-territory on $\mathcal{G}$ with respect to Y if $\boldsymbol{T}$ is closed under descendants and c-component, that is, $De(\boldsymbol{T})_\mathcal{H} = \boldsymbol{T}$ and there is no bidirected edge between $\boldsymbol{T}$ and $\boldsymbol{V} \setminus \boldsymbol{T}$ in $\mathcal{H}$. If there is no UC-territory $\boldsymbol{T}' \subsetneq \boldsymbol{T}$, then $\boldsymbol{T}$ is a minimal UC-territory.* ■

The subgraph induced by Minimal Unobserved-Confounders' Territory (MUCT) represents how Y is determined through the variables ruled by unobserved confounders under the intervention outside the MUCT. MUCT is tightly related to Rule 2 of do-calculus, and the procedure iteratively extends variables that have a backdoor path to Y so that the rest of the variables can be exchangeable between condition and intervention.



(a) $\mathcal{G}$ (b) $\mathcal{G}_{\overline{B}}$ (c) $\text{MUCT}(\mathcal{G}_{\overline{B}}, Y)$ (d) MUCT & $\text{IB}(\mathcal{G}_{\overline{B}}, Y)$

(e) $\mathcal{G}_{\overline{E}}$ (f) MUCT & $\text{IB}(\mathcal{G}_{\overline{E}}, Y)$ (g) $\mathcal{G}_{\overline{F}}$ (h) MUCT & $\text{IB}(\mathcal{G}_{\overline{F}}, Y)$

Figure 32: Obtaining MUCTs (variables in light blue areas) and IBs (variable in green) under intervention (variables in blue) with $\boldsymbol{X}^\star = \boldsymbol{V} \setminus \{Y\}$ (or equivalently $\mathcal{G}$ maybe viewed as the latent projection of an original graph by retaining only $\boldsymbol{X}^\star \cup \{Y\}$). Here, $\mu^*_B \leq \mu^*_D$ and $\mu^*_E \leq \mu^*_{E,F}$.

This has a connection to the partial-orders we are seeking — this demonstrates that $\mu_\emptyset \leq \mu^*_{\boldsymbol{V}\setminus\boldsymbol{T}} = \mu_{pa(\boldsymbol{T})\setminus\boldsymbol{T}}$ if $\boldsymbol{X}^\star$ is defined as $\boldsymbol{V} \setminus \{Y\}$. We define $\text{MUCT}(\mathcal{G}, Y)$ as the MUCT of $\mathcal{G}$ with respect to Y and $\text{IB}(\mathcal{G}, Y) = pa(\boldsymbol{T})_{\mathcal{G}} \setminus \boldsymbol{T}$ as the Interventional Border (IB) of $\mathcal{G}$ with respect to Y where $\boldsymbol{T} = \text{MUCT}(\mathcal{G}, Y)$. For an arbitrary $\boldsymbol{X}^\star \subseteq \boldsymbol{V} \setminus \{Y\}$, one can obtain a MUCT and IB from, say $\mathcal{H}$, the latent projection of $\mathcal{G}$ onto $\boldsymbol{X} \cup \{Y\}$. Then, $\mu^*_{\boldsymbol{W}} \leq \mu^*_{\text{IB}(\mathcal{H}_{\overline{\boldsymbol{W}}}, Y)}$. A more involved example is shown in Figure 32 where MUCT can be procedurally constructed by iteratively updating $\{Y\}$ by including the variables connected with it via bidirected edges and descendants of them. For example, MUCT under $\text{do}(B)$ in Figure 32c can be obtained by first removing A in the graph and expanding $\{Y\}$ with its confounded variables C, its descendants E, and, again, its confounded variable F to ultimately arrive at $\{Y, C, E, F\}$.

Equipped with MUCT and IB, one can check whether there exists a better MIS than a given MIS with respect to their maximum achievable expected rewards. However, the current procedure does not tell us whether there are other MISes better than the one obtained by an interventional border. The following theorem asserts that the interventional border approach provides a way to establishing a complete collection of POMISes given $\mathcal{G}$, $\boldsymbol{X}^\star$, and Y:

Theorem 14 *Given $\mathcal{G}$, $\boldsymbol{X}^\star$, and Y, let $\boldsymbol{X}' \subseteq \boldsymbol{X}^\star$ be an MIS and let $\mathcal{H}$ be the latent projection of $\mathcal{G}$ onto $\boldsymbol{X} \cup \{Y\}$. Then, $\boldsymbol{X}'$ is a POMIS if and only if* $\text{IB}(\mathcal{H}_{\overline{\boldsymbol{X}'}}, Y) = \boldsymbol{X}'$. ■

This result (Lee and Bareinboim, 2018a, 2019a) can be best explained by that, under the distributions of the set of unobserved confounders in a MUCT, the mechanisms of the variables in the MUCT are orchestrated to yield the best result given a configuration (values set outside the MUCT). When any external force (interventions on any variables in the MUCT) is applied, the delicately or-

	do()	do(b)	do(d)	do(e)	do(b, d)	do(d, e)
$P_{d,e}(y \mid a, b)$	$P(y \mid a, b, d, e)$		$P_d(y \mid a, b, e)$	$P_e(y \mid a, b, d)$		
$P_{d,e}(a, b)$	$P(a, b)$		$P_d(a, b)$	$P_e(a, b)$		
$P_{a,c}(e)$	$P(e \mid a, c)$	$P_b(e \mid a, c)$	$P_d(e \mid a, c)$		$P_{b,d}(e \mid a, c)$	
$P_c(d)$	$P(d \mid c)$	$P_b(d \mid c)$		$P_e(d \mid c)$		
$P_b(c)$	$P(c \mid b)$		$P_d(c \mid b)$	$P_e(c \mid b)$	$P_{b,d}(c)$	$P_{d,e}(c \mid b)$

Table 15: For each term shown in Equation 318 (rows), its equal probability quantities that are obtainable from data sampled by playing POMIS arms (columns) are shown. Note, for example, that $P_e(d \mid c) = P_c(d)$ implies that this holds true for every $e \in \mathscr{D}(E)$.

chestrated mechanism is disrupted and a subpar reward is obtained. In Figure 31, arms intervening on $\{X_1\}$, $\{X_3\}$, and $\{X_4\}$ are POMISes.

Property 3. Expressions among actions The two aforementioned structural properties help bandit agents to focus only on a set of minimal arms that can possibly be optimal without examining any collected data. We now consider the third property, which connects the causal effect of playing an arm and the distributions from data acquired through playing other arms.

Consider the causal diagram in Figure 33 with $\boldsymbol{X}^\star = \{B, D, E\}$ where POMISes are $\emptyset$, $\{B\}$, $\{D\}$, $\{E\}$, $\{B, D\}$, and $\{D, E\}$. Then, an online agent playing POMIS arms will obtain samples from $P(\boldsymbol{V})$, $P_b(\boldsymbol{V} \setminus \{B\})$, ..., and $P_{d,e}(\boldsymbol{V} \setminus \{D, E\})$. In this example, one can rewrite the expected reward for, e.g., do($\emptyset$) according to c-factorization (Tian and Pearl, 2002a), as

$$\mathbb{E}[Y] = \sum_{a,b,c,d,e,y} y P_{d,e}(y, a, b) P_{a,c}(e) P_c(d) P_b(c) \tag{317}$$

$$= \sum_{a,b,c,d,e,y} y P_{d,e}(y \mid a, b) P_{d,e}(a, b) P_{a,c}(e) P_c(d) P_b(c) \tag{318}$$

Other arms' expected rewards can be similarly factorized. Here, each term can be replaced by other probabilities obtainable from different POMIS arms with the help of do-calculus (Table 15), which is derived from subsequent applications of do-calculus. These equalities not only imply that a term can be replaced by another but also suggest that each quantity can be estimated from the combination of them. For example, the term $P_c(d)$ can be estimated by a weighted combination of $P_c(d), P(d \mid c), P_e(d \mid c), P_b(d \mid c)$ as

$$\frac{N_c \hat{P}_c(d) + N_{\emptyset|c} \hat{P}(d \mid c) + \sum_e N_{e|c} \hat{P}_e(d \mid c) + \sum_b N_{b|c} \hat{P}_b(d \mid c)}{N_c + N_{\emptyset|c} + \sum_e N_{e|c} + \sum_b N_{b|c}},$$

where $N_{\boldsymbol{w}|\boldsymbol{z}}$ is the number of samples with $\boldsymbol{Z} = \boldsymbol{z}$ in do($\boldsymbol{W} = \boldsymbol{w}$). This expression is nothing but estimating the probability of $D = d$ based on a maximum likelihood principle by aggregating compatible data instances together. Plugging in such estimator for each term in the expression for $\hat{\mathbb{E}}[Y]$ results in an estimator taking advantage of other arms' data.

Imagine an online learning scenario in which $\mu_\emptyset = \mathbb{E}[Y]$ is relatively smaller than the optimal arm. The agent in the scenario will likely play more on other arms with higher rewards. The agent would occasionally play less-played arms when the agent is not completely confident that such arms are not the best arm (e.g., UCB or Thompson sampling). With the expression provided and other

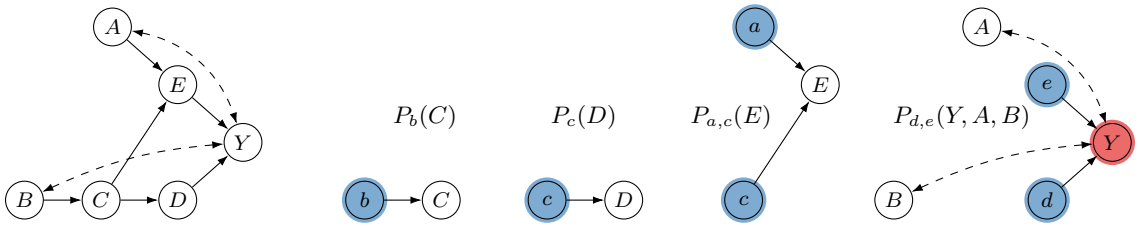


Figure 33: A causal diagram $\mathcal{G}$ and the visualization of original c-factorization of $P(\boldsymbol{v})$.

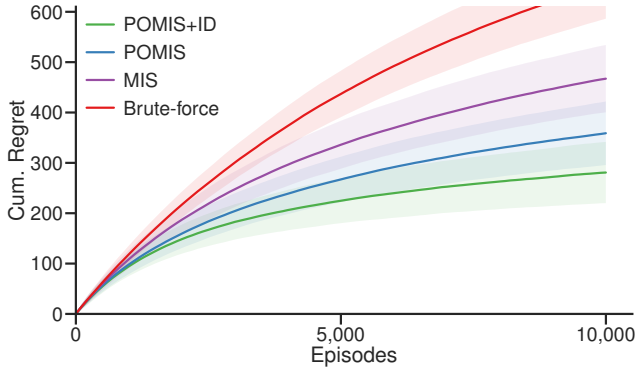
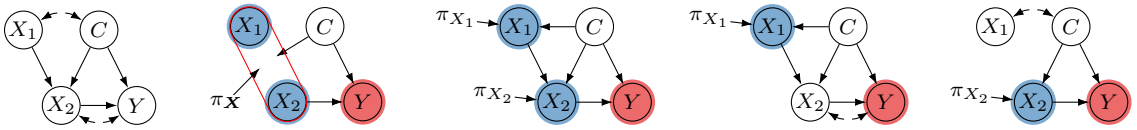


Figure 34: Cumulative regrets of different bandit agents based on Brute-force, MIS, POMIS, and POMIS with identification formula (POMIS+ID). Shaded areas represent standard deviation based on 2000 simulations.

arms' data utilized, the agent can improve confidence on what $\mu_\emptyset$ is by playing other arms only avoiding accumulating regrets.

In (Lee and Bareinboim, 2019a), the posterior of expected reward for each POMIS arm is approximated by bootstrapping samples from multiple data sources and integrated into Thompson sampling. Similarly, based on the variance of expected reward from bootstraps, the effective number of arms played can be approximated and translated back to upper confidence bounds so as to be incorporated into a UCB algorithm.

Experiment 7 *Figure 34 illustrates the cumulative regrets of* `UCB` *agents based on different arm candidates in an environment whose causal diagram is depicted in Figure 33 (leftmost) and Table 15. In this experiment, A and C are not intervenable, and the agent can intervene in the all possible combinations of $\{B, D, E\}$. An agent with brute-force strategy attempts to learn the expected reward for each combination while other agent with MIS or POMIS strategy respectively employs only intervening MISes or POMISes. POMIS+ID indicates an agent actively infers each arm's expected reward from other arms as well if possible. Simulation results illustrate clear gaps among the performances of different strategies. At the end (10,000th episodes), Brute-force, MIS, POMIS, and POMIS+ID respectively yields cumulative regret (mean and standard deviation) of 655.96 ± 69.68, 467.19 ± 66.85, 358.86 ± 63.11, and 280.80 ± 60.83. These results demonstrate that the refinement of arms by considering causal structure improves the efficiency of agents in them interacting with the underlying domain.* ■



(a) $\mathcal{G}$ (b) CB (abstract) (c) $\mathcal{G}_{\Pi_{\text{CB}}}$ (d) $\mathcal{G}_{\Pi_d}$ (e) $\mathcal{G}_{\Pi_e}$

Figure 35: (a) a causal diagram, (b) abstract representation of a contextual bandit policy, and (c,d,e) policy-induced graphs. π nodes are intervention indicators, which will be left implicit throughout the section.

In this section, we presented that an agent should be aware that choosing variables to intervene is not a trivial problem that can be simply answered like, i.e., intervene variables as many as it can, but a sophisticated problem that can be addressed with the knowledge of causal structure. In the next section, we further consider the case where agents can make use of states/contexts in their decision.

6.2 Mixed Policy with Context

As seen in the previous section, a causal understanding of the underlying world enables us to recognize a wide range of policies across various policy spaces, allowing agents to choose their mode of interaction. This involves decisions about not only which variables to intervene in but also to observe as part of the context. In light of this, we explore the use of causal relationships in systematic decision-making over mixed policies with contexts involved. To illustrate the concept of mixed policy with context more clearly, let's consider an agent operating within an environment depicted as in Figure 35a.

Example 47 *In this graph, we have intervenable variables $\boldsymbol{X}^\star = \{X_1, X_2\}$ and contexts $\boldsymbol{S}^\star = \{C, X_1\}$. The primary objective of the agent is to maximize the reward, denoted as μ_π, which is defined as the expected value of Y when following a mixed policy π chosen from a mixed policy space Π_{MIX}.*

In the realm of contextual bandit (CB) problems, the objective is to optimize a policy denoted as π_{CB} (as depicted in Figure 35b). This policy can be viewed as a stochastic mapping from contexts to actions. Alternatively, it can be represented as a pair of decision rules: π_{CB} can be expressed as $(\pi(X_1 \mid C), \pi(X_2 \mid X_1, C))$ or more generally $(\pi(X_1, X_2 \mid C))$ (illustrated in Figure 35c). Traditionally, this policy is optimized within a constrained space denoted as Π_{CB}, which consists of pairs, $\langle X_1, \{C\}\rangle$ and $\langle X_2, \{X_1, C\}\rangle$. However, an issue arises in that the optimal policy π^*_{CB}, determined as $\arg\max_{\pi\in\Pi_{\text{CB}}} \mu_\pi$, may not necessarily be the best possible, i.e., $\mu^*_{\Pi_{\text{CB}}} \triangleq \mu_{\pi^*_{\text{CB}}} < \mu^*$.

Consider a scenario where all variables are binary and U_1 and U_2 are unobserved confounders connected to X_1 and X_2, respectively. Think of these as fair coin flips. Additionally, there's a noise ϵ associated with X_1, which follows a distribution with $P(\epsilon = 1) = 0.2$.

$$\mathscr{F} = \begin{cases} X_1 & \leftarrow U_1 \oplus \epsilon, \\ C & \leftarrow U_1, \\ X_2 & \leftarrow U_2 \oplus X_1 \oplus C, \\ Y & \leftarrow (1 - (X_2 \oplus U_2)) \vee C \end{cases} \tag{319}$$

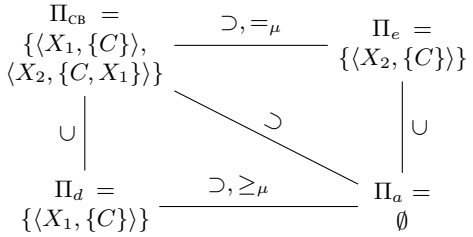


Figure 36: Relationships among the policy spaces based on two aspects.

Given that the chosen policy renders X_2 independent of U_2 and that the context C is also unrelated to U_2, it's deduced that the optimal value for $\mu^*_{\Pi_{\text{CB}}}$ equals 0.75. In this setup, the most effective policy involves intervening solely on X_1 while considering the context C. This policy ensures that, when X_1 is set equal to C, the noise ϵ affecting X_1 is eliminated, resulting in X_2 becoming equivalent to U_2. Consequently, the policy attains an optimal expected reward of 1.0 in this environment. ■

Desiderata for Optimal Mixed Policies From the mixed policy space associated with Figure 35a, we can elicit 15 policy spaces from the mixed policy space. These different modes of interaction can be categorized based on two desiderata: *minimality* and *optimality*. We explain these desiderata through an illustration (Figure 36) of the four policy spaces $\Pi_a = \{\}$, Π_{CB}, $\Pi_d = \{\langle X_1, \{C\}\rangle\}$, and $\Pi_e = \{\langle X_2, \{C\}\rangle\}$ where each subscript represents the label of figure. We say Π subsumes Π', denoted by $\Pi' \subseteq \Pi$, if $\boldsymbol{X}(\Pi') \subseteq \boldsymbol{X}(\Pi)$ and $\boldsymbol{S}'_X \subseteq \boldsymbol{S}_X$, for every $\langle X, \boldsymbol{S}'_X\rangle \in \Pi'$ where $\boldsymbol{X}(\cdot)$ is a set of intervened variables in the policy space. We use $\geq_\mu$ (or $=_\mu$) to indicate whether one's optimal reward is as good as or better than the other's in every scenario compatible with a causal diagram. This establishes equivalence classes among policy spaces based on their optimal rewards.

In simpler terms (to be formalized later on), *minimality* means that removing any actions or contexts from a policy space can worsen its performance. In other words, given two policy spaces Π and Π', if $\Pi \supsetneq \Pi'$ and $\Pi =_\mu \Pi'$, then Π is said to be *redundant*. For instance, since $\Pi_{\text{CB}} \supset \Pi_e$ while $\Pi_{\text{CB}} =_\mu \Pi_e$, the CB policy (Figure 35c) is redundant and the CB agent wastes its resources not only for intervening on X_1 (a redundant action) but also for taking X_1 into account for X_2 (a redundant context).

Furthermore, *optimality* of a policy space Π represents that there exists no other policy space Π' (not in the equivalence class of Π) such that $\Pi' \geq_\mu \Pi$. For example, Π_d, when optimized, is at least as good as Π_a (i.e., $\mu^*_{\Pi_d} \geq \mu^*_{\Pi_a}$) in every environment, and can outperform it in some environments (i.e., $\mu^*_{\Pi_d} > \mu^*_{\Pi_a}$), which demonstrates that Π_a does not meet the optimality criterion. Not all policy spaces can be directly comparable: Π_e is not comparable to Π_a nor Π_d. After careful examination, we find that policy spaces Π_{CB}, Π_d, Π_e meet the optimality criterion. Both minimality and optimality are satisfied only by Π_d and Π_e among all 15 policy spaces. This example illustrates that a smart agent should selectively intervene in variables with relevant contexts to achieve optimal rewards. Against this background, we will delve into the evaluation of mixed policies in terms of their expected rewards.

6.2.1 CONTEXTUAL MINIMALITY IN OPTIMAL MIXED POLICIES

Optimizing a mixed policy involves assessments of the effectiveness of its policy space so that an agent can avoid intervening or observing unnecessary actions or contexts. It is well-known that an

Figure 37: Causal diagrams where the relevance of some contexts can be further eliminated under the optimality of policy.

action on X is worthy if it can affect Y through the change of its mechanism π_X and each context in $S \in \boldsymbol{S}_X$ is relevant to its associated action X if the context provides information relative to other contexts $\boldsymbol{S}_X \setminus \{S\}$ (Lauritzen and Nilsson, 2001; Zhang and Bareinboim, 2020). This simple characterization of (non-)minimality of an individual action and an individual context of policy space is, unfortunately, insufficient to fully grasp, e.g., whether a subset of contexts over multiple actions would be still relevant, especially when $\pi \in \Pi$ is optimized. We explain this insufficiency through an example.

Example 48 *In Figure 37a, both X_1 and X_2 utilize C_3 as their contexts where $\mu_\pi = \mathbb{E}_{C_3}[\mathbb{E}_\pi[Y \mid C_3]]$. Since there exists $c_3^* = \arg\max_{c_3 \in \mathscr{D}(C_3)} \mathbb{E}_\pi[Y \mid c_3]$, we can derive that $\mu_\pi \leq \mathbb{E}_{C_3}[\mathbb{E}_\pi[Y \mid c_3^*]] = \mathbb{E}_\pi[Y \mid c_3^*]$. Given that c_3^* is merely a constant, new decision rules*

$$\pi'(x_i \mid c_1) \triangleq P_\pi(x_i \mid c_i, c_3^*) = \pi(x_i \mid c_i, c_3^*)$$

for $i \in \{1, 2\}$ yield the same optimal reward. That is, X_1 and X_2 (as if they are two agents) agree to assume that C_3 is fixed to some value. ■

This example first demonstrates the idea of *fixing* where it is viable to treat a variable in the context as a fixed value without sacrificing optimal performance. Once fixed, decision rules in the mixed policy can free the variable from being a context, resulting in a simpler policy space. A more sophisticated example is shown in Figure 37b where a redundant context can be *fixed conditionally* on the remaining contexts.

Example 49 *In the policy illustrated in Figure 37b, both intervened variables are relying on the same contexts C_1 and C_2. The expected reward is expressed as*

$$\mu_\pi = \sum_{y, \boldsymbol{x}, \boldsymbol{s}} y P_{\boldsymbol{x}}(y, c_1, c_2) \pi(x_1 \mid c_1, c_2) \pi(x_2 \mid c_1, c_2).$$

Given that $P_{\boldsymbol{x}}(y, c_1, c_2) = P_{\boldsymbol{x}}(c_2 \mid y, c_1) P_{\boldsymbol{x}}(y, c_1) = P(c_2 \mid c_1) P_{\boldsymbol{x}}(y, c_1)$ based on basic do-calculus,

$$= \sum_{c_1, c_2} P(c_2 \mid c_1) \underbrace{\sum_{y, \boldsymbol{x}} y P_{\boldsymbol{x}}(y, c_1) \pi(x_1 \mid c_1, c_2) \pi(x_2 \mid c_1, c_2)}_{\text{define } \mu_\pi(c_1, c_2)} \tag{320}$$

This expression can be viewed as the sum of weighted rewards for each C_1 value. Let c_2^* be a function taking c_1 such that $c_2^*(c_1) = \arg\max_{c_2} \mu_\pi(c_1, c_2)$ for $c_1 \in \mathscr{D}(C_1)$. Then,

$$\begin{aligned}
&\leq \sum_{c_1,c_2} P(c_2 \mid c_1)\mu_\pi(c_1, c_2^*(c_1)) \\
&= \sum_{c_1} \mu_\pi(c_1, c_2^*(c_1)) \\
&= \sum_{y,c_1,\boldsymbol{x}} yP_{\boldsymbol{x}}(y, c_1)\pi(x_1 \mid c_1, c_2^*(c_1))\pi(x_2 \mid c_1, c_2^*(c_1)).
\end{aligned}$$

By incorporating c_2^* into π, we can devise a smaller mixed policy π' such that $\pi'(x_1 \mid c_1) = \pi(x_1 \mid c_1, c_2^*(c_1))$ and $\pi'(x_2 \mid c_1) = \pi(x_2 \mid c_1, c_2^*(c_1))$.

$$= \sum_{y,c_1,\boldsymbol{x}} yP_{\boldsymbol{x}}(y, c_1)\pi'(x_1 \mid c_1)\pi'(x_2 \mid c_1) = \mu_{\pi'}.$$

■

The key idea demonstrated in this derivation is finding contexts that can be fixed to some desirable values relative to other contexts (or none) so that decision rules can be equivalently performed without relying on the contexts fixable via remaining contexts. Relationships between the two types of contexts best captured in Equation 320. More generally, the values to be fixed does not have to be contexts to yield an expression that is 'greater than equal to' the previous expression. The restriction is (to meet our purpose to rewrite the expression to the expected reward of a smaller policy space) that any fixed context should be inferred from the rest of context (e.g., c_2 relative to c_1). Further, this process may involve the use of intervened variables that are fixed (determined) by their contexts under optimality (to show later).

Against this background, we will define and characterize the minimality of policy space under optimality, which has practical implications to an agent learning an optimal policy.

Definition 20 (Minimality under Optimality) *Given $\langle \mathcal{G}, Y, \boldsymbol{X}^\star, \boldsymbol{S}^\star \rangle$, a policy space Π is said to be* minimal under optimality *if there exists an SCM $\mathcal{M}$ compatible with $\mathcal{G}$ such that $\mu_\Pi^* > \mu_{\Pi'}^*$ for every strictly subsumed policy space $\Pi' \subsetneq \Pi$, that is,*

$$\exists \mathcal{M} \sim \mathcal{G}\, \forall \Pi' \subsetneq \Pi\, (\mu_\Pi^* > \mu_{\Pi'}^*).$$

■

We will develop a sufficient condition for non-minimality under optimality by generalizing the idea presented earlier. The condition is made of two parts. The first part is obtaining a specific form of an intermediate expression for expected reward given a set of variables to fix. In the next part, based on the intermediate expression, it checks whether the variables can indeed be fixed to yield a new simpler policy. Before we proceed to investigate conditions for finding a simpler policy, let us briefly discuss what we mean by *fixing*. Consider the following form of expression,

$$\sum_{a,\boldsymbol{b},\boldsymbol{d}} P(a \mid \boldsymbol{b}, \boldsymbol{c}) f(a, \boldsymbol{b}, \boldsymbol{d}) \leq \sum_{\boldsymbol{b},\boldsymbol{d}} f(a^*(\boldsymbol{b}), \boldsymbol{b}, \boldsymbol{d}) = \sum_{\boldsymbol{b},\boldsymbol{d}} f'(\boldsymbol{b}, \boldsymbol{d}).$$

With $\boldsymbol{C}$ fixed to a constant, we say a is *fixed conditional* on $\boldsymbol{B}$. For the purpose of eliciting a simpler policy, decision rules (implicit in f) also drop a from its argument by inferring it from $\boldsymbol{b}$ and $\boldsymbol{c}$.

Step 1: obtaining an intermediate expression from an expected reward Given a policy space Π satisfying the basic minimality, let $\boldsymbol{X}' \subseteq \boldsymbol{X}(\Pi)$ be actions of interest (of which we would like to change its decision rules), $\boldsymbol{S}' \subsetneq \boldsymbol{S}_{\boldsymbol{X}'} \setminus \boldsymbol{X}'$ non-action contexts of interest (that is, contexts to keep among the contexts of $\boldsymbol{X}'$). Given (i) a subset of exogenous variables $\boldsymbol{U}'$ in $\mathcal{G}_\Pi$[49], (ii) a subset of endogenous variables $\boldsymbol{Z}$ in $\mathcal{G}_\Pi$ that disjoints with $\boldsymbol{S}' \cup \boldsymbol{X}'$ and subsumes unselected contexts $\boldsymbol{S}_{\boldsymbol{X}'} \setminus (\boldsymbol{S}' \cup \boldsymbol{X}')$, and (iii) an order $\prec$ over $\boldsymbol{V}' \triangleq \boldsymbol{S}' \cup \boldsymbol{X}' \cup \boldsymbol{Z}$, if the triple $\langle \boldsymbol{U}', \boldsymbol{Z}, \prec \rangle$ satisfies certain conditions (Lee and Bareinboim, 2020, Lemma 1), then we can write μ_π as

$$\mu_\pi = \overbrace{\sum_{\boldsymbol{u}'} P_\pi(\boldsymbol{u}')}^{\text{marginally fixable}} \sum_{y, \boldsymbol{s}', \boldsymbol{x}'} y \underbrace{Q'_{\boldsymbol{x}'}(y, \boldsymbol{s}')}_{\text{irrelevant to } \boldsymbol{Z}} \overbrace{\sum_{\boldsymbol{z}} \underbrace{\prod_{Z \in \boldsymbol{Z}} P_\pi(z \mid \boldsymbol{v}'_{\prec Z}, \boldsymbol{u}')}_{\text{defines dependency}}}^{\text{to fix conditionally}} \prod_{X \in \boldsymbol{X}'} \underbrace{\pi(x \mid \overbrace{\boldsymbol{s}_X \setminus \boldsymbol{z}}^{\text{given}}, \overbrace{\boldsymbol{s}_X \cap \boldsymbol{z}}^{\text{to infer}})}_{\text{to become } \pi'(x \mid \boldsymbol{s}_X \setminus \boldsymbol{z})}, \tag{321}$$

where $Q' = P_{\pi \setminus \boldsymbol{X}'}$ is a distribution under π except $\boldsymbol{X}'$. The conditions are mainly designed to elicit the term $Q'_{\boldsymbol{x}'}(y, \boldsymbol{s}')$ without $\boldsymbol{Z}$ so that fixing does not affect the term, and later we can transform the expression into the expected reward for a simpler policy relying only on context $\boldsymbol{S}'$.

We explain the intermediate expression. To begin with we replace $P_\pi(Z \mid \boldsymbol{V}'_{\prec Z}, \boldsymbol{U}')$ to $P_\pi(Z \mid \boldsymbol{V}_Z)$ where $\boldsymbol{V}_Z$ is a minimal subset of $\boldsymbol{V}'_{\prec Z} \cup \boldsymbol{U}'$ dependent to Z. The purpose of intermediate expression is to be transformed to $\mu_{\pi'}$ such that all decision rules $(\pi(X \mid \boldsymbol{S}_X))$ for $X \in \boldsymbol{X}'$ become $(\pi(X \mid \boldsymbol{S}_X \setminus \boldsymbol{Z}))$. We achieve this by fixing all the contexts containing $\boldsymbol{Z}$ as dictated in $P_\pi(z \mid \boldsymbol{v}_Z)$ similar to $P(c_2 \mid c_1)$ in the earlier example. What we have found is that fixing variables other than unnecessary contexts can ultimately help to fix and remove those unnecessary contexts. This explains why some of the exogenous variables and variables other than unnecessary contexts are involved. Finally, the order explicitly decides how probability terms are factorized following a chain rule and, thus, how variables are fixed relative to other variables.

Step 2: transforming intermediate expression into the expected reward for a simpler policy Once we attain the intermediate expression for some $\langle \boldsymbol{U}', \boldsymbol{Z}, \prec \rangle$, we examine whether the expression can be converted to $\mu_{\pi'}$ where $\Pi' \triangleq (\Pi \setminus \boldsymbol{X}') \cup \{\langle X, \boldsymbol{S}_X \setminus \boldsymbol{Z} \rangle\}_{X \in \boldsymbol{X}'}$. Algorithmically speaking, we can first fix $\boldsymbol{u}'$ unconditionally. Then, check whether any currently unfixed $Z \in \boldsymbol{Z}$ can be fixed since $\boldsymbol{V}'_{\prec Z}$ are all fixed, which are made of contexts, actions, and other endogenous variables. Other than fixing value of Z to the best possible value to drop the term $P(Z \mid \cdot)$, action variables can be fixed in a similar way.

Example 50 *Through Figure 38, we will show that, C_2 and C_3 are* redundant contexts under optimality. *Given $\boldsymbol{S}' = \{C_1\}$ and $\boldsymbol{X}' = \{X_1, X_2\}$, consider $\boldsymbol{Z} = \{C_2, C_3\}$, $\boldsymbol{U}' = \emptyset$, and order $\prec = \langle C_3, C_1, X_2, C_2, X_1 \rangle$. We can derive the following expression for the expected reward (with subscripts concatenated),*

$$\begin{aligned}
\mu^*_\Pi &= \sum_{y, \boldsymbol{x}, c_1} y Q'_{\boldsymbol{x}}(y \mid c_1) \sum_{c_{23}} P_\pi(c_{123}, \boldsymbol{x}) \\
&= \sum_{y, \boldsymbol{x}, c_1} y Q'_{\boldsymbol{x}}(y \mid c_1) \sum_{c_{23}} P_\pi(c_3) P_\pi(c_1 \mid c_3) P_\pi(x_2 \mid c_{13}) P_\pi(c_2 \mid c_{13}, x_2) P_\pi(x_1 \mid c_{123}, x_2)
\end{aligned}$$

49. Formally speaking, we are selecting unobserved variables associated with a clique formed by bidirected edges in $\mathcal{G}_\Pi$.

(a) (b)

Figure 38: (a) A minimal policy space and (b) its dependency graph derived in Example 50 where C_1 is given.

$$= \sum_{c_3} P_\pi(c_3) \sum_{y,\boldsymbol{x},c_1} yQ'_{\boldsymbol{x}}(y,c_1) \sum_{c_2} P_\pi(c_2 \mid c_{13}, x_2)\pi(x_2 \mid c_3)\pi(x_1 \mid c_{12}). \tag{322}$$

C_3 can be fixed to a constant c_3^* so that,

$$\leq \sum_{y,\boldsymbol{x},c_1} yQ'_{\boldsymbol{x}}(y,c_1) \sum_{c_2} P_\pi(c_2 \mid c_1, c_3^*, x_2)\pi(x_2 \mid c_3^*)\pi(x_1 \mid c_1, c_2).$$

This expression can be rearranged so that $\sum_{x_2} \pi(x_2 \mid c_3^*)$ starts the expression where there exists $x_2^* \in \mathscr{D}(X_2)$, which allows us to substitute $\pi(x_2 \mid c_3^*)$ with $\pi'(x_2)$ where $\pi'(x_2^*) = 1$.

$$\leq \sum_{y,\boldsymbol{x},c_1} yQ'_{\boldsymbol{x}}(y,c_1) \sum_{c_2} P_\pi(c_2 \mid c_1, c_3^*, x_2^*)\pi'(x_2)\pi(x_1 \mid c_1, c_2).$$

Here, although we can drop both x_2 from summation and $\pi'(x_2)$ from the expression, we keep them to connect to the expected reward of the resulting simpler policy. Next, the optimal c_2 is determined with respect to c_1, i.e., $P_\pi(c_2 \mid c_1, c_3^*, x_2^*)$, where we can replace $\pi(x_1 \mid c_1, c_2^*(c_1))$ by $\pi'(x_1 \mid c_1)$. We start by reordering terms for readability.

$$\begin{aligned}
&= \sum_{c_1,c_2} P_\pi(c_2 \mid c_1, c_3^*, x_2^*) \sum_{y,\boldsymbol{x}} yQ'_{\boldsymbol{x}}(y,c_1)\pi'(x_2)\pi(x_1 \mid c_1, c_2) \\
&\leq \sum_{y,\boldsymbol{x},c_1} yQ'_{\boldsymbol{x}}(y,c_1)\pi'(x_2)\pi(x_1 \mid c_1, c_2^*(c_1)) \\
&= \sum_{y,\boldsymbol{x},c_1} yQ'_{\boldsymbol{x}}(y,c_1)\pi'(x_1 \mid c_1)\pi'(x_2) = \mu^*_{\Pi'}.
\end{aligned} \tag{323}$$

Since $\mu^*_{\Pi'} \leq \mu^*_{\Pi}$ by the existence of $\pi \in \Pi$ that can emulate $\pi' \in \Pi'$, and $\mu^*_{\Pi'} \geq \mu^*_{\Pi}$ by the derivation (Equation 323), we can conclude that $\mu^*_{\Pi'} = \mu^*_{\Pi}$. As a consequence, policy space Π is not minimal under optimality due to the ineffective contexts $\{C_2, C_3\}$ with respect to $\{X_1, X_2\}$. ■

The procedure leading to a simpler policy from the intermediate expression can be described as constructing and examining a dependency graph as follows. Based on the distributions over $\boldsymbol{Z}$ and $\boldsymbol{U}'$, we can construct a dependency graph in the form of DAG. Initially, vertices are $\boldsymbol{U}'$, $\boldsymbol{Z}$, $\boldsymbol{Z}$'s parents, and the parents of intervened variables. Directed edges are added so that the parents of each node is the set of conditions in its distribution, that is, $P_\pi(Z \mid \boldsymbol{V}_Z)$ yields Z having $\boldsymbol{V}_Z$

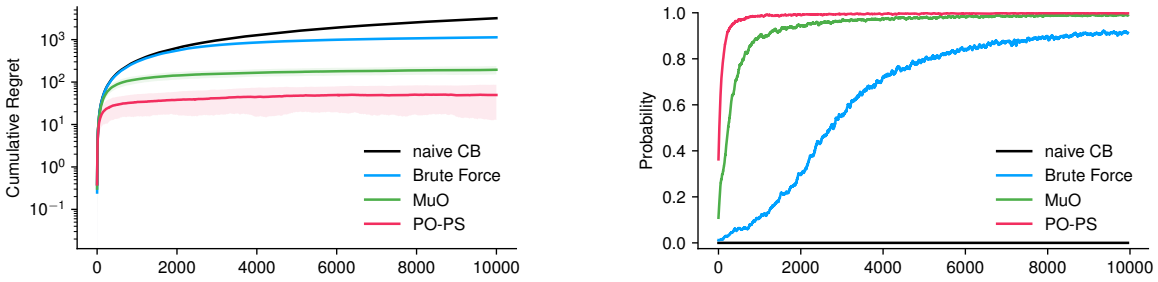


Figure 39: (left) Cumulative regrets (in a log scale) of different arm strategies based on all possible mixed policy spaces (brute-force, BF), naive contextual bandit (CB), Minimality under Optimality (MuO), and possibly optimal policy spaces (PO-PS). Shaded areas represent standard deviation based on 100 simulations. (right) Probability each agent selecting the optimal arm.

as its parents in the graph. This similarly applies to the decision rules $\pi(x \mid \boldsymbol{s}_x)$. The dependency graph for the Example 50 is shown in Figure 38b with $\boldsymbol{Z} = \{C_2, C_3\}$ and their parents determined through factorization in Equation 322.

Once a dependency graph is constructed, we can then figure out whether all $\boldsymbol{Z}$'s can be fixed to yield a simpler policy. Contexts to keep $\boldsymbol{S}'$ do not have parents in the dependency graph and are marked to indicate they are available to fix other variables. Here, C_1 is available to fix, e.g., C_2, for its values. Nodes are marked if its parents are all marked. For example, C_3 can be marked unconditionally. X_2 is marked given that C_3 is marked. With C_3 and X_2 marked, altogether with C_1, C_2 is marked, resulting in fixing to $c_2^*(c_3^*, x_2^*(c_3^*), c_1) = c_2^*(c_1)$. Finally, if (i) all $\boldsymbol{Z}$'s in the dependency graph are marked, and (ii) the ancestors of each intervened variable X do not include any of context variables that will not be available in the simpler policy, that is, $\boldsymbol{S}' \setminus (\boldsymbol{S}_X \setminus \boldsymbol{Z})$, then we can yield a simpler policy. More detailed results are presented in Theorem 2 (Lee and Bareinboim, 2020).

Experiment 8 *Figure 39 depicts the cumulative regrets of* `UCB` *agents based on different arm selection strategies in the environment described in Example 47. There are four strategies where each strategy considers a subset of policy spaces in the mixed policy space. Brute-force makes use of all policy spaces. A naive contextual bandit (CB) implements intervening both given C (Figure 35c), which can be modeled as intervening on X_1 given $\{C\}$ and X_2 given $\{C, X_1\}$. The minimality under optimality (MuO) agent plays only the arms without redundant interventions or contexts. Finally, possibly optimal policy spaces (PO-PS) discards policy spaces among the minimal policy spaces that are not strictly better than other policy spaces (see Lee and Bareinboim (2020) for details). With a smaller number of policy spaces to consider, PO-PS converges faster than other strategies. MuO is better than the brute-force agent, which finds out the optimal policy among all possible options. Although brute-force agent is slow, it converges since the optimal policy space is included. However, the naive CB agent is unable to converge since, in this example, intervening on X_2 leads to suboptimal policy due to the mechanism involving the unobserved confounder between X_2 and Y. Average cumulative regret of PO-PS, MuO, BF, and naive CB respectively was 49.60, 195.03, 1134.84, and 3198.42.* ■

Discussion Mixed policy learning provides a flexible framework when variables to be intervened associated with natural mechanisms for how their values to be determined. We have showed that

naively intervening on the underlying system while ignoring natural mechanisms can result in a suboptimal policy, incurring regrets indefinitely. We have characterized mixed policies with and without contexts. In the case without contexts, complete characterizations for the minimality and possible-optimality of a policy space within a given mixed policy space are provided. With context, we provide a procedure to detect non-minimality. These characterizations do not require samples through interacting the given environment. Lee and Bareinboim (2020) provide accounts on the possible-optimality of policy space with context. A complete characterization of both minimality and possible-optimality (partial-order) for the case of mixed policy space with contexts is an open problem.

7. Counterfactual Decision-Making (CRL Task 3)

In this section, we investigate a novel type of interaction between the agent and the environment via layer 3 of the PCH. For the tasks described so far, the agent interacts with the underlying environment through passive observations (seeing), active experimentation (doing), or a combination of both, which in the context of the PCH evokes layers 1 and 2-interactions. The new interactive modality will allow the agent to search over the large space of counterfactual policies, Π_{CTF}. More specifically, an online learning task with counterfactual randomization is characterized by a signature defined as follows:

$$\mathcal{T}_{\text{ctf-rand}} = \langle \mathcal{I} = \text{ctf}, \mathcal{A} = \emptyset, \Pi = \Pi_{\text{CTF}}, \mathcal{R} = \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R} \rangle \,. \tag{324}$$

This means that the agent will try to find a policy π^* such that

$$\pi^* = \arg\max_{\pi \in \Pi_{\text{CTF}}} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R}\left(\boldsymbol{Y}\right) \;\middle|\; \mathcal{D}_{\text{ctf}} \sim P\left(\boldsymbol{V}_{\boldsymbol{x}} \mid \boldsymbol{x}'\right) \right], \tag{325}$$

where the distinct feature of the task is the counterfactual type of interactions. To make this argument more precise, recall that for an experimental policy space Π_{EXP} (Def. 8), performing an intervention $\text{do}(\pi)$ following a policy $\pi \in \Pi_{\text{EXP}}$ induces an interventional distribution $P_{\pi}\left(\boldsymbol{V}\right)$ (Def. 5), which follows from Fisherian randomization. On the other hand, interaction following a counterfactual policy $\pi \in \Pi_{\text{CTF}}$ allows the agent to access a specific type of counterfactual distribution, the x-specific effect of the decision on the outcome (Plecko and Bareinboim, 2022, Sec. 4.1.1).[50]

As it will become clear throughout this section, we will introduce a new type of counterfactual randomization (for short, *ctf-rand*), which will allow the agent to behave optimally in challenging decision-making settings. The following example illustrates one of such scenario using an instance of MAB environment, graphically described in Fig. 10a.

Example 51 (Greedy Casino) *A group of investors decide to develop a new casino in Las Vegas and wants to make their machines as lucrative as possible at all costs, which we will call the Greedy casino incorporated (GCI). GCI's owners are determined about their mission and divide their efforts in three phases: research, setup, and operations.*

50. This quantity is also known in the literature as the effect of treated on the treated (ETT). For further discussion and historical context, refer to (Heckman, 1992) and (Pearl, 2000, Ch. 8.2.).

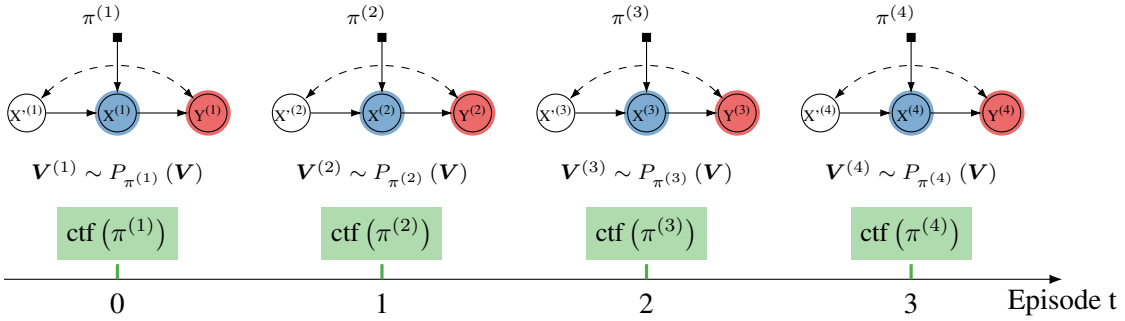


Figure 40: Temporal diagram showing an agent interacting with the environment for repeated episodes through counterfactual policies.

Phase 1. Research *GCI's executives hire a team of cognitive scientists, psychologists, and cognitive scientists to investigate human's behaviors in the casinos' floors currently in operation around town. The team conducts a battery of studies and discovers that two features, out of hundreds examined, accurately predict the gambling behavior of players on a casino floor: each player's inebriation and the machine's conspicuousness (e.g., whether a machine is blinking and making noise). Coding these traits as binary variables, we let* $U_B \in \{0, 1\}$ *denote whether or not a machine is blinking, and* $U_D \in \{0, 1\}$ *denote whether or not the gambler is drunk.*

As another outcome of the team's comprehensive study, they discover that the gambling population tends to prefer attracting less attention and naturally tends to be shy. However, their behavior changes when they become intoxicated, and they are more drawn towards the more effusive machines, such as those blinking and making noise. Formally, a gambler's 'natural' choice is described by the following mechanism (starting at 0):

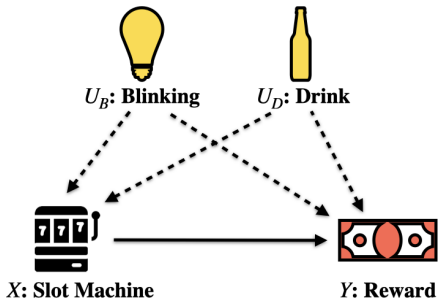
$$X \leftarrow f_X(U_B, U_D) = \neg(U_D \oplus U_B), \tag{326}$$

where $X = 1$ *represents staying in the current machine, and* $X = 0$ *represents switching to the neighbor machine. For instance, the gambler will stay in the slot machine they are currently in ("*$X = 1$*") whenever the current machine is blinking (*$U_B = 1$*) and he is drunk (*$U_D = 1$*), or this machine is not blinking (*$U_B = 0$*) and he is sober (*$U_D = 0$*). Alternatively, the gambler will get uncomfortable and switch machines ("*$X = 0$*") whenever the machine is blinking (*$U_B = 1$*) but he is not drunk (*$U_D = 0$*), or the machine is not blinking (*$U_B = 0$*) and he is drunk (*$U_D = 1$*). The table in Fig. 16b summarizes this behavior.*

Phase 2. Setup *The GC owners take the information gathered during the research stage and devise a new plan to leverage it in order to maximize profitability. Specifically, they purchase brand new machines with several important capabilities, including:*

1. *High-definition cameras capable of recording the gamblers' faces and body language.*
2. *New lighting and sound systems that enable the machines to blink and make noise.*
3. *The latest deep learning software that allows the machines to analyze the gamblers' behavior.*

The machines can be configured to operate in a network, so that they can share capabilities to make the look and feel of the casino's floor more pleasant. In practice, GC's executives decide to have the machines operate in pairs, meaning they share the same source of randomness. This config-



(a)

U_B	U_D	X
0	0	1
0	1	0
1	0	0
1	1	1

(b)

Figure 41: (a) Illustration of the machines configurations. (b) Table summarizing gamblers natural predispositions.

	$U_D = 0$		$U_D = 1$	
	$U_B = 0$	$U_B = 1$	$U_B = 0$	$U_B = 1$
$X = 0$	0.50	*0.10	*0.20	0.40
$X = 1$	*0.10	0.50	0.40	*0.20

(a)

	$\mathbb{E}[Y\|X]$	$\mathbb{E}_x[Y]$
$X = 0$	0.15	0.3
$X = 1$	0.15	0.3

(b)

Table 16: (a) Payout rates $P(Y = 1 \mid X, U_D, U_B)$ decided by reactive slot machines as a function of arm choice, sobriety, and machine conspicuousness. Players' natural arm choices under U_D, U_B are indicated by asterisks. (b) Payout rates according to the observational, $P(Y = 1|X)$, and interventional $P_x(Y = 1)$, distributions, where $Y = 1$ represents winning (shown in the table).

uration is illustrated in Fig. 41a, where a blinking/noisy machine is paired with a non-blinking/silent one.

The gambler arrives at one of the machines and can immediately play it or switch to the alternative one. (Importantly, the gamblers are already drunk or not just before arriving, and the machines are already blinking or not regardless of the gambler, those are two independent events.) We will call the former by machine 1 ($X = 1$) and the latter by 0 ($X = 0$). Also, the outcome of each gamble is represented by a variable Y, where 0 means that the patron lost, and 1 otherwise.

Moreover, to keep the system in an equilibrium, and under the radar, the casino distributes free drinks and sets up the machines such that every gambler has an equal chance of being intoxicated and each machine has an equal chance of blinking its lights at a given time. In formal notation, this means that $P(U_D = 0) = P(U_D = 1) = 0.5$ and $P(U_B = 0) = P(U_B = 1) = 0.5$.

The owners are also cognizant of current's state gambling regulations that require casinos to maintain a minimum attainable payout rate for slots of 30%. While still wanting to maximize profits, GCI executives decide to take advantage of the players' propensities by leveraging the machines' sensing capabilities. They then set up the payout rates f_Y of the machines as depicted in Table 16a. In words, if $U_D = 0, U_B = 0$, the player will naturally select action $X = 1$, following Eq. 326, which will lead to a positive outcome, $Y = 1$, only 10% of the time (the same with $U_D = 1, U_B =$

1*). Also, if* $U_D = 0, U_B = 1$ *(or* $U_D = 1, U_B = 0$*), the player will decide for action* $X = 0$*, to switch, which will lead to a positive outcome* 20% *of the time. These configurations are marked with an* $*$ *in the table.*

Phase 3. Operational *The GC debuts and is a big hit; many new patrons enjoy their evenings playing in the new machines. Interestingly, they are not aware (conscious) that their behavior is influenced by their inebriation and whether the machine is looking conspicuous. The variables* U_B, U_D *are exogenous and remain unobserved, following the causal language introduced earlier.*

Still, some patrons know about GC owners' reputations and are suspicious of the casino's ethical standards. These patrons decide to collect some data on the other gamblers' behavior, through random sampling, which leads to the distribution shown in Table 16b. In other words, it seems that the casino is paying ordinary gamblers only 15% of the time. Also, no matter whether they play machine $X = 0$ *or* $X = 1$*, the average payout is the same.*

The state is called to investigate the issue and, being blind to the GC's payout strategy, claims that this data is observational (non-causal), and, therefore, is "invalid". They then decide to conduct a randomized study to verify whether the win rates in the floor meet the legal standards. The government's inspectors follow the RCT procedure discussed in Sec. 4.2. First, they recruit random players from the casino floor, pay them to play a random slot, and then observe the outcome. The experiment yields a favorable outcome for the casino, with win rates precisely meeting the 30% cutoff – no more, no less than. The data looks like Table 16b, and is again insensitive to the machine's choice.

As RL enthusiasts, we decide to run a series of experiments using more refined and sample-efficient adaptive strategies (e.g., ϵ*-greedy, Thompson Sampling, UCB1, EXP3) to test the new slot machines on the casino's floor. We obtain data encoded in Fig. 42. The first plot shows that the probability of choosing the correct action is no better than a random coin flip even after a considerable number of steps. We note, somewhat surprised, that the cumulative regret continues without abating, indicating our inability to learn a superior arm. We also realize that the results obtained by the standard algorithms align with the randomized study (orange line).*

After all, the casino seems to be, at the same time, (1) exploiting gamblers' natural predilections as a function of their intoxication and the machine's blinking behavior (based on Eq. 326), (2) paying, on average, less than the legally allowed (15% instead of 30%), and (3) fooling state's inspectors since the randomized trial payout meets the legal requirement. ■

Some observations are worth noting after this example. Firstly, the situation described in the greedy casino is far from contrived. There is a growing body of literature in the cognitive sciences that recognizes a significant aspect of human decision-making occurring at a subconscious level, with individuals often unaware of the reasons behind their actions (Kouider et al., 2010).[51]

Second, under the presence of unobserved confounders, such as in the GC example, the interventional quantity $\mathbb{E}_x[Y]$ used throughout the RL literature does not seem to capture critical information required to maximize payout, but rather the average payout akin to choosing arms by a coin flip (as shown in the plot earlier). Specifically, the payout given by coin flipping is the same

51. Interestingly, the work of psychology Professor Daniel Kahneman, a Nobel Prize laureate, revolves around recognizing and studying various biases and mechanisms in human decision-making. For more insights, refer to (Tversky and Kahneman, 1974; Bargh and Chartrand, 1999; Dijksterhuis and Nordgren, 2006) for a survey on these results.

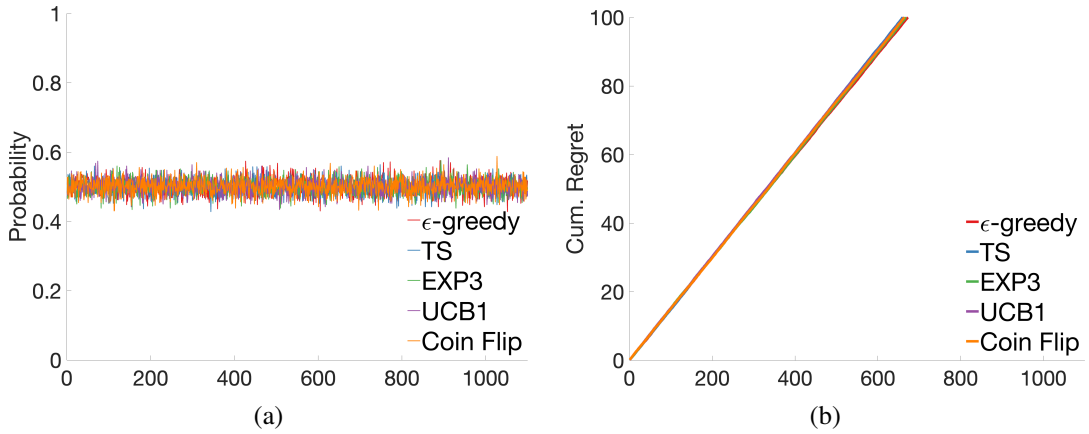


Figure 42: Performance of different bandit strategies in the greedy casino example; x-axis represents the total episodes of interactions. (a) No algorithm is able to perform better than random guessing. (b) Regret grows without bounds.

for both machines,

$$\mathbb{E}_{X\leftarrow 0}[Y] = \mathbb{E}_{X\leftarrow 1}[Y] = 0.3, \tag{327}$$

which means that the arms are statistically indistinguishable in the limit of a large sample size. Further, if we consider using the observational data from watching gamblers on the casino floor (based on their natural predilections), the average payoff is also independent of the machine choice,

$$\mathbb{E}[Y = 1 \mid X = 0] = \mathbb{E}[Y = 1 \mid X = 1] = 0.15, \tag{328}$$

albeit with an even lower payout. Based on these observations, we can see why no arm choice is better than the other under either distribution alone, which explains the reason any algorithm based on these distributions will fail to learn an optimal policy.

Third, and more fundamentally, one may be puzzled by the discrepancy between observational and interventional distributions. After all, even though the interventional distribution refers to the causal effect, the typical player receives payouts that come from the observational data; how can this be reconciled? Furthermore, could this difference reveal insights about the unobserved confounders, offering clues on how to differentiate the arms? Lastly, from a more practical standpoint, acknowledging this phenomenon and considering the data in Table 16b, what would be the optimal way to play at the GC? Is it possible to devise a strategy that yields a higher payout than the two methods previously discussed?

In this section, our goal is to further understand and answer these questions. To achieve this, we will introduce novel causal machinery designed to exploit PCH's layer 3 distributions, thereby enabling the agent to achieve higher performance in challenging decision-making scenarios, such as the greedy casino discussed in Example 51. We aim to formalize the concept of counterfactual randomization, including canonical environments such as MABs (Fig. 10a) and MDPs (Fig. 10d). We will provide a systematic augmentation procedure that empowers existing online learning agents to optimize counterfactual policies in these environments. The contributions of these sections are summarized as follows:

- Sec. 7.1 proposes a novel *counterfactual decision criterion* in bandit models. This strategy determines the values of actions based on a specific type of counterfactual quantity, namely, the effect of the treatment on the treated. This approach leads to a new family of counterfactual policies that take advantage of the agent's intended actions (i.e., intuitions). Sec. 7.1.1 extends this counterfactual criterion to a canonical family of sequential decision-making environments, specifically Markov decision processes with unobserved confounders (MDP).

- Sec. 7.2 develops a novel *counterfactual randomization* to enable the realization of counterfactual decision-making in the underlying environment. Sec. 7.2.1 extends this novel randomization procedure to enhance existing online algorithms in MDP environments. We demonstrate this procedure by augmentating a state-of-art MDP algorithm `UCBVI` (Azar et al., 2017). The analysis suggests that the new counterfactual randomization strategy is statistically efficient and consistently outperforms standard online algorithms that lack counterfactual reasoning.

- Sec. 7.3 connects the idea of counterfactual policies and the notion of autonomy in decision-making. In particular, we introduce a novel trade-off between *autonomy and optimality*: while a fully autonomous system is preferable, discarding the human input could harm the optimal performance of the decision system, leading to suboptimal strategies. An effective planning algorithm is developed to balance this autonomy-optimality trade-off, enabling an agent to learn an optimal counterfactual policy in an MDP environment under a budget constraint on the frequency of using human input.

7.1 Counterfactual Decision Criterion

We begin the discussion by describing the causal mechanisms that encode the agent's interaction regimes with the MAB environment. First, when the agent passively observes events unfolding in the underlying causal model, the underlying causal mechanisms remain unchanged. Fig. 43a illustrates the causal diagram of a canonical MAB environment. Note that f_X represents the system's behavioral policy generating the observational data. It takes as input the unobserved factors U affecting the reward signal Y and determines an observed arm choice X; the presence of U is represented by the bidirected arrow $X \dashleftarrow\dashrightarrow Y$ (Def. 6). One way to interpret this arrow is through the concept of players' natural predilections. For instance, in the greedy casino (Example 51), the predilection could correspond to choices made by gamblers when allowed to play freely on the casino's floor (e.g., intoxicated players favoring blinking machines), or doctors prescribing drugs based on their "gut feeling" (e.g., physicians prescribing the more expensive drug to wealthier patients). The rewards associated with these predilections are encoded in the observational quantity $\mathbb{E}[Y \mid x]$.

On the other hand, the interventional quantity $\mathbb{E}_x[Y]$ encodes the reward induced by the process in which the natural predilections are overridden or ceased by external and deliberate policies. In the casino's example, this reward arises when the government's inspectors flip a coin and direct gamblers to machines based on the coin's outcome via intervention $\text{do}(x)$, regardless of their natural predilections. Fig. 43b depicts the causal diagram of the post-interventional MAB environment. Here, the bidirected arrow between X and Y is removed, as the unobserved confounder U does not influence the intervention $\text{do}(x)$ and how the value of X attains its value. The exogenous

distribution $P(U)$ and the reward function $f_Y(X, U)$ encompass the expected reward parameters for each arm, which are typically the focus of analysis in RL literature.[52]

Among the two interaction regimes described above, the observational agent ("see") does not deliberately search for a better alternative other than its natural predilections. Meanwhile, the interventional agent ("do") explores the environment by randomly playing arms, thus disregarding its natural predilections. Remarkably, it is possible to leverage the information embedded in these distinct interaction regimes (and their corresponding distributions) to understand and improve the agents' natural predilections in these MAB instances.

To witness, assume the agent is now more introspective and starts operating through the following protocol on the casino's floor. The gambler observes itself and intercepts its decision-flow when is just about to pull the arm of machine "0". He then contemplates whether following his natural predilection ("0") or going against it (playing "1") would lead to a better outcome. The drunk gambler, for example, who is a clever machine learning student and familiar with Figs. 43a and 43b, says that such evaluation cannot be computed a priori. He affirms that, despite spending hours on the casino estimating the expected payoff based on players' natural predilections (namely, $\mathbb{E}[Y \mid x]$), it is not feasible to relate this natural predilection with the hypothetical construction what would have happened had he decided to play differently. He further acknowledges that the interventional quantity $\mathbb{E}_x[Y]$, devoid of the gamblers' predilections, does not support any clear comparison against his personal strategy. The oracle says this type of reasoning is possible, but first one needs to define the concept of counterfactual distributions:

Definition 21 (Counterfactual Distribution (Bareinboim et al., 2020)) *An SCM $\mathcal{M} = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P \rangle$ induces a family of joint distributions over counterfactual events $\boldsymbol{Y}_{\boldsymbol{x}}, \ldots, \boldsymbol{Z}_{\boldsymbol{w}}$, for any $\boldsymbol{Y}, \boldsymbol{Z}, \ldots, \boldsymbol{X}, \boldsymbol{W} \subseteq \boldsymbol{V}$:*

$$P(\boldsymbol{y}_{\boldsymbol{x}}, \ldots, \boldsymbol{z}_{\boldsymbol{w}}) \equiv \sum_{\boldsymbol{u}} \mathbb{1}\{\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u}) = \boldsymbol{y}, \ldots, \boldsymbol{Z}_{\boldsymbol{w}}(\boldsymbol{u}) = \boldsymbol{w}\} P(\boldsymbol{u}), \tag{329}$$

■

Note that the l.h.s. of Eq. 329 contains variables with different subscripts, which, syntactically, encode different counterfactual worlds. The evaluation implied by this equation can be described as the following process:

1. For each set of subscripts relative to each set of variables (e.g., $\text{do}(\boldsymbol{x}), \ldots, \text{do}(\boldsymbol{w})$ for $\boldsymbol{Y}, \ldots, \boldsymbol{Z}$, respectively), replace the corresponding mechanisms with the appropriate constants and generate $\mathscr{F}_{\boldsymbol{x}}, \ldots, \mathscr{F}_{\boldsymbol{w}}$ (Def. 3), creating submodels $\mathcal{M}_{\boldsymbol{x}}, \ldots, \mathcal{M}_{\boldsymbol{w}}$ (of $\mathcal{M}$);

2. For each situation $\boldsymbol{U} = \boldsymbol{u}$, the environment evaluates the modified causal mechanisms (e.g., $\mathscr{F}_{\boldsymbol{x}}, \ldots, \mathscr{F}_{\boldsymbol{w}}$) following a valid order (i.e., any variable in the l.h.s. is evaluated after the ones in the r.h.s.) to obtain the potential responses of the observables, and

3. The probability mass $P(\boldsymbol{U} = \boldsymbol{u})$ is then accumulated for each instantiation $\boldsymbol{U} = \boldsymbol{u}$ that is consistent with the events over the counterfactual variables – for instance, $\boldsymbol{Y}_{\boldsymbol{x}} = \boldsymbol{y} \ldots, \boldsymbol{Z}_{\boldsymbol{w}} = \boldsymbol{z}$, i.e., $\boldsymbol{Y} = \boldsymbol{y}, \ldots, \boldsymbol{Z} = \boldsymbol{z}$ in the submodels $\mathcal{M}_{\boldsymbol{x}}, \ldots, \mathcal{M}_{\boldsymbol{w}}$, respectively.

52. The standard MAB literature focuses on the unconfounded case depicted in Fig. 43b), where $\mathbb{E}_x[Y] = \mathbb{E}[Y \mid x]$ (Def. 13). Throughout this section, we focus on the general setting where the NUC assumption does not hold and the unobserved confounder U cannot be ruled out a priori.

(a) See (b) Do (c) Ctf

Figure 43: Causal diagrams for different interaction regimes in the MAB environment.

Among quantities in Def. 21, the counterfactual event $\boldsymbol{Y}_{\boldsymbol{x}}(\boldsymbol{u}) = \boldsymbol{y}$ could be read as a sentence: "$\boldsymbol{Y}$ would be $\boldsymbol{y}$ (in situation $\boldsymbol{U} = \boldsymbol{u}$), had $\boldsymbol{X}$ been $\boldsymbol{x}$." This definition of counterfactuals naturally leads to the judgment suggested by the oracle.

Example 52 (Counterfactual reasoning through SCMs (GC continued)) *Consider again the MAB model $\mathcal{M}^*$ describing the greedy casino environment in Example 51. We are interested in evaluating whether the new statement, "Would I (the agent) win ($Y = 1$) had I played $X = 0$?" can be formally written as a counterfactual event. Assuming that the agent's natural predilection is to play machine $X = 1$, the agent now engages in introspection to compare the odds of winning following his "gut feeling" or going against his intuition. This statement can be written in counterfactual language, formally, as*

$$\mathbb{E}\left[Y_{X\leftarrow 0} \mid X = 1\right], \tag{330}$$

which reads as "the expected value of winning ($Y = 1$) had I played $X = 0$ given that I am about to play $X = 1$". This statement contrasts with the alternative quantity

$$\mathbb{E}\left[Y_{X\leftarrow 1} \mid X = 1\right] \tag{331}$$

which reads as "the expected value of winning ($Y = 1$) had I played $X = 1$ given that I am about to play $X = 1$". This quantity is also called the X-specific effect of X on Y (Plecko and Bareinboim, 2022, Sec. 4.1.1)..

More specifically, for quantities in the form of Eq. 331, the composition axiom (Pearl, 2000, Ch. 7.3) implies that

$$\mathbb{E}\left[Y_{X\leftarrow 1} \mid X = 1\right] = \mathbb{E}\left[Y \mid X = 1\right], \tag{332}$$

where the l.h.s. is computable from the observational distribution $P(X, Y)$. Using the previous discussion in Example 51, computing the above equation gives

$$\mathbb{E}\left[Y_{X\leftarrow 1} \mid X = 1\right] = 0.15 \tag{333}$$

The counterfactual quantity in the form of Eq. 330 could be written as

$$\mathbb{E}\left[Y_{X\leftarrow 0} \mid X = 1\right] = P(Y_{X\leftarrow 0} = 1 \mid X = 1) \tag{334}$$

$$= \frac{P(Y_{X\leftarrow 0} = 1, X = 1)}{P(X = 1)} \tag{335}$$

	$\mathbb{E}[Y_x \mid X=0]$	$\mathbb{E}[Y_x \mid X=1]$
$X=0$	0.15	*0.45
$X=1$	*0.45	0.15

Table 17: Payout rates according to the counterfactual distribution $P(Y_{X\leftarrow x} \mid X = x')$.

where the denominator is trivially obtainable since it only involves observational probabilities. Following the behavioral policy in Eq. 326, we obtain

$$P(X=1) = P(U_D=0, U_B=1) + P(U_D=1, U_B=0) \tag{336}$$
$$= 0.5 \tag{337}$$

On the other hand, the numerator, $P(Y_{X\leftarrow 1}=1, X=0)$ is more interesting and refers to two different worlds and cannot be written in the languages of observational and interventional distributions since they do not allow for probability expressions involving more than one subscript (each encoding a different version of the environment). Using the procedure dictated in Eq. 329, we obtain

$$P(Y_{X\leftarrow 0}=1, X=1) = \sum_{u_B,u_D} \mathbb{1}\{Y_{X\leftarrow 0}(u_B,u_D)=1, X(u_B,u_D)=1\}P(u_B,u_D) \tag{338}$$

Noting that $U_D=1, U_B=1$ and $U_D=0, U_B=0$ are not compatible with the event $X=0$, through Eq. 326, we can write:

$$P(Y_{X\leftarrow 0}=1, X=1) = 0.25 \times \Big(\{Y_{X\leftarrow 0}(1,0)=1, X(1,0)=1\} \tag{339}$$
$$+ \{Y_{X\leftarrow 0}(0,1)=1, X(0,1)=1\}\Big) \tag{340}$$

Considering the first factor and going back to f_Y, as shown in Table 12a, we note that $Y=1$ when $D=0, B=1, X=0$ with probability 0.5, and when $D=1, B=0, X=0$ with probability 0.4. Putting this back into Eq. 335 leads to:

$$P(Y_{X\leftarrow 0}=1, X=1) = \frac{0.25}{0.5} * 0.9 = 0.45 \tag{341}$$

For completeness, we also compute the other values for the x-specific effect, $\mathbb{E}[Y_x \mid x']$, for all $x, x' = 0, 1$, as shown in Table 17. The conclusion following this analysis is clear – the payout rate would have tripled had the agent played machine $X=0$ in situations where its natural predilections suggest $X=1$, and machine $X=1$ in situations where its natural predilections suggest $X=0$. ■

The counterfactual analysis in the previous example suggests a novel decision-making criterion in the MAB environment. Instead of using a decision rule comparing the average payouts associated with arms, namely (for action x),

$$x^* = \arg\max_x \mathbb{E}_x[Y], \tag{342}$$

we should consider the rule using the comparison between the average payouts obtained by players for choosing in favor or against their intuition, respectively,

$$\pi^*(x) = \arg\max_x \mathbb{E}\left[\underbrace{Y_{X\leftarrow x}}_{\text{realized action}} \mid \underbrace{X = x'}_{\text{intended action}}\right], \quad \forall x \in \{0, 1\} \tag{343}$$

where x' is the player's natural predilection, and x is their final decision.

We call this procedure *counterfactual decision criterion* (CDC) to emphasize the counterfactual nature of this reasoning step and the idea of either following or disobeying the agent's intuition. Remarkably, CDC takes into account the agent's individuality and the fact that their natural inclination provides valuable information about the confounders that also affect the payout, importantly, even when unknown to the very agent. In the binary case, for example, assuming that $X = 1$ is the player's natural choice at some time step, if

$$\mathbb{E}\left[Y_{X\leftarrow 0} \mid X = 1\right] \geq \mathbb{E}\left[Y_{X\leftarrow 1} \mid X = 1\right], \tag{344}$$

this would suggest that the player should act against their intuition, i.e., refraining of playing machine $X = 1$ in favor of playing machine $X = 0$. Conversely, if

$$\mathbb{E}\left[Y_{X\leftarrow 0} \mid X = 1\right] \leq \mathbb{E}\left[Y_{X\leftarrow 1} \mid X = 1\right], \tag{345}$$

it would imply that the player should follow their intuition, in this case, playing machine $X = 1$. Performing CDC leads to a novel type of counterfactual policies in MAB environment.

Definition 22 (Counterfactual Policy - MAB) *Let $\mathcal{M}$ be an MAB environment graphically described in Fig. 43a. A counterfactual policy space Π_{CTF} is a collection of policies $\pi(X \mid X')$ mapping from the domain of an intended action X' to the space of probability distribution over the domain of a realized action X. Henceforth, we will consistently denote such a policy space by $\Pi_{\text{CTF}} = \{\langle X, \{X'\}\rangle\}$.*

At first glance, a counterfactual policy $\pi(X \mid X')$ might be counter-intuitive since the input and output of the policy π appear to be the same, over the domain of action variable X. This suggests that deploying a counterfactual policy introduces a self-reference in the underlying environment. Semantically speaking, this is a well-defined counterfactual quantity and computable for any SCM $\mathcal{M}$. While its input X' and output X share the same domain, they represent action variables in two different worlds – the input X' is the agent's natural predilection, similar to the obtained in the observational distribution in layer 1 (Fig. 43a), while the output X is the agent's realized action generating the interventional distribution in layer 2 (Fig. 43b). Determining the realized intervention $\text{do}(X \leftarrow x)$ based on the agent's natural predilection $X = x'$ leads to the expected reward of a counterfactual policy $\pi(X \mid X')$.

Definition 23 (Counterfactual Submodel, MAB) *Let $\mathcal{M}$ be an MAB environment graphically described in Fig. 43a such that*

$$\mathcal{M} = \langle \boldsymbol{U} = \{U\}, \boldsymbol{V} = \{X, Y\}, \mathscr{F} = \{f_X, f_Y\}, P = P(U)\rangle \tag{346}$$

Let $\pi(X \mid X')$ be a counterfactual policy over action X. A submodel $\mathcal{M}_\pi$ of $\mathcal{M}$ is a modified SCM

$$\mathcal{M}_\pi = \left\langle \boldsymbol{U} = \{U\}, \boldsymbol{V}_\pi = \{X', X, Y\}, \mathscr{F}_\pi, P = P(U)\right\rangle, \tag{347}$$

where the structural functions $\mathscr{F}_\pi$ are defined as

$$\mathscr{F}_\pi = \begin{cases} X' \leftarrow f_X(U) \\ X \sim \pi(X \mid X') \\ Y \leftarrow f_Y(X', U) \end{cases} \tag{348}$$

■

Fig. 43c provides a graphical representation of the submodel $\mathcal{M}_\pi$ induced by a counterfactual policy $\pi(X \mid X')$. Note that a new, virtual variable X' is added to represent the intended action, which is used as an input to determine the realized action X affecting the reward signal Y. Formally, we define the expected reward $\mathbb{E}_\pi[Y; \mathcal{M}]$ of a counterfactual policy $\pi(X \mid X')$ evaluated in an MAB environment as the expected reward $\mathbb{E}[Y; \mathcal{M}_\pi]$ evaluated in the submodel $\mathcal{M}_\pi$. That is,

$$\mathbb{E}_\pi[Y; \mathcal{M}] = \sum_{x', x} \mathbb{E}\left[Y \mid x, x'; \mathcal{M}_\pi\right] P\left(x \mid x'; \mathcal{M}_\pi\right) P\left(x'; \mathcal{M}_\pi\right) \tag{349}$$

$$= \sum_{x', x} \mathbb{E}\left[Y \mid x, x'; \mathcal{M}_\pi\right] \pi\left(x \mid x'\right) P\left(x'\right) \tag{350}$$

The last step holds since in the counterfactual submodel $\mathcal{M}_\pi$ defined by Eq. 348, the values of the intended action X' are decided by the behavioral policy f_X determining action in the original SCM $\mathcal{M}$. Since the counterfactual policy π does not take the unobserved confounder U affecting other variables in the system, conditioning on the intended action X' "blocks" the backdoor paths from the realized action X to the subsequent reward Y. Computing the expected reward of Y conditioning on both the intended and realized actions X', X in submodel thus recovers the counterfactual ETT quantity. The above equation can be further written as

$$\mathbb{E}_\pi[Y; \mathcal{M}] = \sum_{x, x'} \mathbb{E}\left[Y_{X \leftarrow x} \mid X = x'\right] \pi\left(x \mid x'\right) P\left(x'\right) \tag{351}$$

The following example demonstrates counterfactual policies in the Greedy Casino environment described previously in Example 51.

Example 53 *Consider the MAB environment $\mathcal{M}^*$ described in Example 51. Performing intervention ctf(π) following a counterfactual policy $\pi \triangleq X \leftarrow \neg X'$ defines a submodel $\mathcal{M}^*_\pi$ described by the following tuple*

$$\mathcal{M}^*_\pi = \left\langle \boldsymbol{U} = \{U_D, U_B\}, \boldsymbol{V}_\pi = \{X', X, Y\}, \mathscr{F}_\pi, P = P(U_D, U_B) \right\rangle \tag{352}$$

The structural functions in $\mathscr{F}_\pi$ are defined as

$$\mathscr{F}_\pi = \begin{cases} X' \leftarrow U_D \oplus U_B \\ X \leftarrow \neg X' \\ Y \leftarrow f_Y(X, U_D, U_B) \end{cases} \tag{353}$$

Table 18 shows the detailed parametrization of the joint distribution $P(X, X', Y)$ evaluated in the

	$X'=0$		$X'=1$	
	$X=0$	$X=1$	$X=0$	$X=1$
$Y=0$	0	0.275	0.275	0
$Y=1$	0	0.225	0.225	0

Table 18: The joint distribution $P(X', X, Y)$ evaluated in the submodel $\mathcal{M}^*_\pi$ induced by a counterfactual policy $\pi \triangleq X \leftarrow \neg X'$.

counterfactual submodel $\mathcal{M}^*_\pi$ described in Example 53. The expected reward of Y in submodel $\mathcal{M}^*_\pi$ is given by

$$\mathbb{E}_\pi[Y] = P\left(X'=0, X=0, Y=1\right) + P\left(X'=1, X=0, Y=1\right) \tag{354}$$
$$+ P\left(X'=0, X=1, Y=1\right) + P\left(X'=1, X=1, Y=1\right) \tag{355}$$

Computing the above equation gives $\mathbb{E}_\pi[Y] = 0.45$, which outperforms the expected reward of atomic intervention $\mathbb{E}_x[Y] = 0.15$ for every arm $x = 0, 1$. ■

More generally, note that the counterfactual policy space Π_{CTF} contains the experimental policy $\Pi_{\text{EXP}} = \{\langle X, \emptyset\rangle\}$ in MAB environments. For every experimental policy $\pi(X)$, one could simulate it using a counterfactual $\pi(X \mid X)$ by selecting realized action $\text{do}(X \leftarrow x)$ regardless of natural predilection $X' = x'$, i.e., $\pi(x) = \pi(x \mid x')$, $\forall x, x'$. In this case, the expected reward in Eq. 351 could be further written as:

$$\mathbb{E}_\pi[Y] = \sum_x \pi(x) \sum_{x'} \mathbb{E}\left[Y_{X\leftarrow x} \mid X = x'\right] P(x') \tag{356}$$
$$= \sum_x \pi(x) \mathbb{E}[Y_x] \tag{357}$$

By the definition of potential outcomes (Def. 4) and interventional distributions (Def. 5), the counterfactual quantity $\mathbb{E}[Y_x] = \mathbb{E}_x[Y]$. The above equation thus coincides with the expected reward of an experimental policy $\pi(x)$. As shown next, an optimal counterfactual policy consistently dominates the best possible experimental policy in terms of performance.

Theorem 15 (Counterfactual dominates Interventional Policies (MAB)) *For an MAB environment $\mathcal{M}^*$, let policy spaces $\Pi_{\text{CTF}} = \{\langle X, \{X\}\rangle\}$ and $\Pi_{\text{EXP}} = \{\langle X, \emptyset\rangle\}$. Then, an optimal counterfactual policy is never worse than an optimal interventional policy, namely,*

$$\arg\max_{\pi\in\Pi_{\text{CTF}}} \mathbb{E}_\pi[Y] \geq \arg\max_{\pi\in\Pi_{\text{EXP}}} \mathbb{E}_\pi[Y] \tag{358}$$

■

A natural question at this point is when the equality in the equation holds, and the standard interventional agent is able to achieve the optimal performance of an counterfactual agent. When the NUC condition (Def. 13) holds in the underlying MAB environment, there is no unobserved confounder

affecting the action X and the reward Y, simultaneously. This implies that the agent's intended action X is independent of the potential outcome Y_x induced by intervention do(x), namely,[53]

$$(X \perp\!\!\!\perp Y_x) \tag{359}$$

When the above independence relationship holds, Eq. 351 could be further written as:

$$\mathbb{E}\left[Y; \mathcal{M}_\pi\right] = \sum_x \mathbb{E}\left[Y_{X \leftarrow x}\right] \sum_{x'} \pi\left(x \mid x'\right) P\left(x'\right) \tag{360}$$

$$= \sum_x \mathbb{E}\left[Y_{X \leftarrow x}\right] P_\pi(x) \tag{361}$$

In the last step, probabilities $P_\pi(x) = \sum_x \pi\left(x \mid x'\right) P\left(x'\right)$ are obtained by marginalizing over the domain of the intended action X. An agent could thus simulate the performance of the counterfactual policy $\pi(x'|x)$ using an experimental policy $\pi(x) = P_\pi(x)$. In other words, the optimal performance of counterfactual and experimental policies coincide whenever the NUC holds.

7.1.1 MARKOV DECISION PROCESS WITH UNOBSERVED CONFOUNDERS

The remainder of this section expands on the concept of counterfactual policies to a more general sequential decision-making setting where the agent must decide on the values of a sequence of actions. Our discussion will focus on a canonical family of environments that extend Markov decision processes (Puterman, 1994) through the language of structural causality.

Definition 24 (MDP Environment) *Consider an SCM describing an MDP environment*

$$\mathcal{M} = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P(\boldsymbol{U}) \rangle, \tag{362}$$

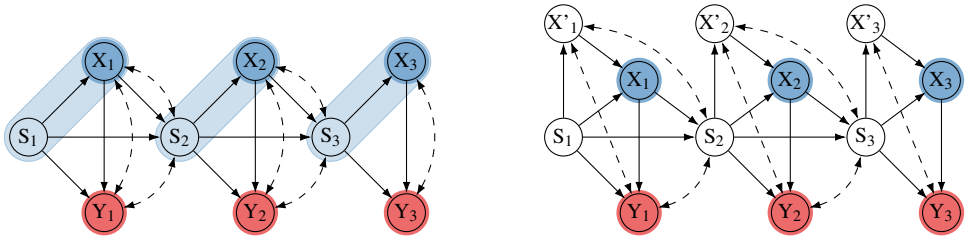
where for a decision horizon $H \in \mathbb{N}^+$,[54]

- $\boldsymbol{U} = \{U_1, \ldots, U_H\}$ *is a sequence of exogenous variables U_i;*
- $\boldsymbol{V} = \{\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{Y}\}$ *is a set of endogenous variables consisting of a sequence of states $\boldsymbol{S} = \{S_1, \ldots, S_H\}$, actions $\boldsymbol{X} = \{X_1, \ldots, X_H\}$, and rewards $\boldsymbol{Y} = \{Y_1, \ldots, Y_H\}$;*
- $\mathscr{F}$ *is a set of functions determining values of $\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{Y}$ such that for every $i = 1, \ldots, H$,*[55]

$$\mathscr{F} = \begin{cases} S_i \leftarrow f_S(S_{i-1}, X_{i-1}, U_i) \\ X_i \leftarrow f_X(S_i, U_i) \\ Y_i \leftarrow f_Y(S_i, X_i, U_i) \end{cases} \tag{363}$$

- P *is a joint distribution over $\boldsymbol{U}$ such that $P(\boldsymbol{U}) = \prod_{i=1}^{H} P(U_i)$ and $U_1, \ldots, U_H$ are i.i.d. variables drawn over a domain $\mathscr{D}(U)$, i.e., $P(U_1) = \cdots = P(U_H)$.*[56]

53. This independence relationship is also referred to as *ignorability* in the literature (Rosenbaum and Rubin, 1983).
54. The decision horizon H could be finitely large, i.e. $H = \infty$. The model $\mathcal{M}$ is called an infinite-horizon MDP.
55. With a slight abuse of notation, we denote by $S_{i-1} = \emptyset$, $X_{i-1} = \emptyset$ if $i = 1$, i.e., the initial state $S_1 \leftarrow f_S(U_1)$.
56. Compared with dynamic treatment regimes (Murphy et al., 2001a), MDPs explicitly encode the *locality* in both the underlying causal mechanisms and exogenous noises affecting states, actions and rewards at every time step. This structural constraint manifests in the Markov property in system dynamics, as discussed in Sec. 3.3.



(a) (b)

Figure 44: Causal diagrams for the MDP environment and its submodel induced by a counterfactual policy $\pi = (\pi_1(X'_1 \mid S_1, X_1), \ldots, \pi_H(X'_H \mid S_H, X_H))$.

■

Def. 24 describes a generalized family of environments similar to MDPs, where the NUC condition does not hold and unobserved confounders are not excluded a priori. Consequently, this family of environments is referred to in the literature as MDP with unobserved confounders (MDPUCs) (Zhang and Bareinboim, 2022). In an MDP environment, the Markov property holds for both the observational distribution $P(\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{Y})$ and interventional distribution $P_\pi(\boldsymbol{S}, \boldsymbol{X}, \boldsymbol{Y})$ induced by a policy $\pi = (\pi_1(X_1 \mid S_1), \ldots, \pi_H(X_H \mid S_H))$. For every time step $i = 1, \ldots, H$, all the future state $\bar{\boldsymbol{S}}_{i+1:H}$, actions $\bar{\boldsymbol{X}}_{i:H}$ and rewards $\bar{\boldsymbol{Y}}_{i:H}$ are independent of the history $\bar{\boldsymbol{S}}_{1:i-1}, \bar{\boldsymbol{X}}_{1:i-1}, \bar{\boldsymbol{Y}}_{1:i-1}$ given the current state S_i, namely:

$$\left(\bar{\boldsymbol{S}}_{i+1:H}, \bar{\boldsymbol{X}}_{i:H}, \bar{\boldsymbol{Y}}_{i:H} \perp\!\!\!\perp \bar{\boldsymbol{S}}_{1:i-1}, \bar{\boldsymbol{X}}_{1:i-1}, \bar{\boldsymbol{Y}}_{1:i-1} \mid S_i\right) \tag{364}$$

The above independence relationship could be read from the causal diagram of the MDP environment, shown in Fig. 44a, following the d-separation rules (Def. 7). However, due to the presence of unobserved confounders, the transition probabilities and conditional reward in the observational and interventional distributions do not necessarily coincide, i.e. for every step $i = 1, \ldots, H$,

$$P(s_{i+1} \mid s_i, x_i) \neq P_{x_i}(s_{i+1} \mid s_i) \tag{365}$$

$$\mathbb{E}[Y_i \mid s_i, x_i] \neq \mathbb{E}_{x_i}[Y_i \mid s_i] \tag{366}$$

Recall the more detailed discussion of the Markov property and the NUC assumption in Sec. 3.3. For instance, Example 2 shows an MDP environment concerning the inventory management of a retail store; potential unobserved confounders include human errors of the store manager, uncertainties in the customers' demands, and monetary values of the goods. Both its observational and interventional distributions can be compactly represented using finite-state automata. However, detailed parameters of these automata differ significantly due to the presence of unobserved confounders; detailed computation is provided in Examples 23 and 24.

We next formalize the concept of counterfactual policies in MDP environments. Similar to MABs, an agent following a counterfactual policy can be thought of as selecting the values of every action $X_i \in \boldsymbol{X}$ based on its original intended action. However, unlike in the previous MAB settings, the agent will also consider observed values of the current S_i before taking action X_i.

Definition 25 (Counterfactual Policy - MDP) *For an MDP environment $\mathcal{M}^*$, a counterfactual policy space Π_{CTF} is a collection of policies*

$$\pi = \left(\pi_1(X_1 \mid S_1, X'_1), \ldots, \pi_H(X_H \mid S_H, X'_H)\right), \tag{367}$$

where every decision rule $\pi_i(X_i \mid S_i, X'_i)$ is a function mapping from the domain of state S_i and intended action X'_i to the space of probability distribution over the domain of realized action X_i. Henceforth, we will consistently denote such a policy space by $\Pi_{\text{CTF}} = \{\langle X_i, \{S_i, X'_i\}\rangle\}_{i=1}^{H}$. ■

Similar to MAB environments, an agent interacting an MDP environment $\mathcal{M}$ following a counterfactual policy π leads to a submodel $\mathcal{M}_\pi$ with additional intended actions X'_i mediating between every realized action X_i and its direct parents, including the current state S_i and the unobserved confounder U_i. Formally,

Definition 26 (Counterfactual Submodel, MDP) *Let $\mathcal{M} = \langle \boldsymbol{U}, \boldsymbol{V}, \mathscr{F}, P(\boldsymbol{U})\rangle$ be an MDP environment, and $\pi = (\pi_1(X_1 \mid S_1, X'_1), \dots, \pi_H(X_H \mid S_H, X'_H))$ be a counterfactual policy over actions $X_1, X_2, \dots$. A submodel $\mathcal{M}_\pi$ of $\mathcal{M}$ is an SCM*

$$\mathcal{M}_\pi = \left\langle \boldsymbol{U}, \boldsymbol{V}_\pi = \left\{\boldsymbol{S}, \boldsymbol{X}', \boldsymbol{X}, \boldsymbol{Y}\right\}, \mathscr{F}_\pi, P = P(\boldsymbol{U})\right\rangle, \tag{368}$$

where $\boldsymbol{X} = \{X_1, \dots, X_H\}$ is a sequence of realized actions; $\mathscr{F}_\pi$ is a set of structural functions defined as

$$\mathscr{F}_\pi = \begin{cases} S_i \leftarrow f_S(S_{i-1}, X_{i-1}, U_i) \\ X'_i \leftarrow f_X(S_i, U_i) \\ X_i \sim \pi_i(X_i \mid S_i, X'_i) \\ Y_i \leftarrow f_Y(S_i, X_i, U_i) \end{cases} \tag{369}$$

■

The causal diagram in Fig. 44b is associated with the submodel $\mathcal{M}_\pi$ induced by an MDP environment and a counterfactual policy $\pi(X_i, \mid S_i, X'_i)$. Formally, we define $P_\pi\left(\boldsymbol{S}, \boldsymbol{X}', \boldsymbol{X}, \boldsymbol{Y}\right)$ of a counterfactual policy $\pi(X_i, \mid S_i, X'_i)$ as the joint distribution over endogenous variables $\boldsymbol{S}, \boldsymbol{X}', \boldsymbol{X}, \boldsymbol{Y}$ in submodel $\mathcal{M}_\pi$. One could see by inspection that the data-generating mechanisms in Fig. 44b define a Markov chain (Puterman, 1994). For every stage of intervention $i = 1, 2, \dots$, the state S_i and intended action X'_i satisfy the Markov property with regard to past state and actions' history. More specifically, the following independent relationships hold in the counterfactual submodel $\mathcal{M}_\pi$,

$$P\left(S_{i+1}, X'_{i+1} \mid \bar{\boldsymbol{S}}_{1:i}, \bar{\boldsymbol{X}}'_{1:i}, \bar{\boldsymbol{X}}_{1:i}; \mathcal{M}_\pi\right) = P\left(S_{i+1}, X'_{i+1} \mid S_i, X'_i, X_i; \mathcal{M}_\pi\right) \tag{370}$$

$$\mathbb{E}\left[Y_i \mid \bar{\boldsymbol{S}}_{1:i}, \bar{\boldsymbol{X}}'_{1:i}, \bar{\boldsymbol{X}}_{1:i}; \mathcal{M}_\pi\right] = \mathbb{E}\left[Y_i \mid S_i, X'_i, X_i; \mathcal{M}_\pi\right] \tag{371}$$

The following example demonstrates the Markov property in a counterfactual MDP submodel. The following example demonstrates counterfactual policies in MDP environments.

Example 54 (Autonomous vs. Semi-autonomous Systems) *Consider the MDP environment $\mathcal{M}^*$ described in Eq. 5, where the decision horizon $H = \infty$. Recall that the experimental policy space $\Pi_{\text{EXP}} = \{\langle X_i, \{S_i\}\rangle\}_{i=1}^{\infty}$ contains a collection of policies $\pi = (\pi_1(X_1 \mid S_1), \pi_2(X_2 \mid S_2), \dots)$. Every decision rule $\pi_i(X_i \mid S_i)$ is a probability distribution mapping from state S_i to action X_i. Operationally, the experimental decision model $\langle \mathcal{M}^*, \Pi_{\text{EXP}}, \mathcal{R}\rangle$ defines an autonomous inventory management system that determines whether to refill X_i based on the current inventory size S_i. Note that the manager's intended decision X_i is not accounted in the system's decision-making process, and can thus be discarded.*

S_{i+1}	X_{i+1}	S_i	X'_i	X_i	P	S_{i+1}	X_{i+1}	S_i	X'_i	X_i	P
0	0	0	0	0	0.09	1	0	0	0	0	0.09
0	0	0	0	1	0.01	1	0	0	0	1	0.81
0	0	0	1	0	0.01	1	0	0	1	0	0.81
0	0	0	1	1	0.09	1	0	0	1	1	0.09
0	0	1	0	0	0.09	1	0	1	0	0	0.09
0	0	1	0	1	0.09	1	0	1	0	1	0.09
0	0	1	1	0	0.01	1	0	1	1	0	0.81
0	0	1	1	1	0.01	1	0	1	1	1	0.81
0	1	0	0	0	0.09	1	1	0	0	0	0.01
0	1	0	0	1	0.01	1	1	0	0	1	0.09
0	1	0	1	0	0.01	1	1	0	1	0	0.09
0	1	0	1	1	0.09	1	1	0	1	1	0.01
0	1	1	0	0	0.09	1	1	1	0	0	0.01
0	1	1	0	1	0.09	1	1	1	0	1	0.01
0	1	1	1	0	0.01	1	1	1	1	0	0.09
0	1	1	1	1	0.01	1	1	1	1	1	0.09

Table 19: Evaluation of the counterfactual transition distribution $P(S_{i+1_{x_i}}, X_{i+1_{x_i}} \mid S_i, X_i = x'_i)$ evaluated in the MDP environment of Example 2.

We now consider the counterfactual policy space $\Pi_{\text{CTF}} = \{\langle X_i, \{S_i, X_i\}\rangle\}_{i=1}^{\infty}$. *Every counterfactual policy* $\pi \in \Pi_{\text{CTF}}$ *is a sequence of decision rules* $(\pi_1(X'_1 \mid S_1, X_1), \pi_2(X'_2 \mid S_2, X_2), \dots)$. *Operationally, the counterfactual decision model* $\langle \mathcal{M}^*, \Pi_{\text{CTF}}, \mathcal{R}\rangle$ *defines an inventory management system that repeatedly calibrates the manager's intended action* X_i *based on the observed state* S_i. *Note that this decision-making system is semi-autonomous since it proactively accounts for the manager's decision. Compared with the autonomous system described above, the counterfactual intervention does not entirely replace the human behavioral policy operating in the environment.*[57]

More specifically, let a counterfactual policy $\pi = (\pi_1(X_1 \mid X_1, S_1), \pi_2(X_2 \mid X_2, S_2), \dots)$ *such that for every* $i = 1, 2, \dots$, *the decision rule* π_i *is defined as,*

$$\pi_i \triangleq X_i \leftarrow S_i \oplus X'_i \oplus U_{i,4} \tag{372}$$

where $U_{i,4}$ *is a new independent noise uniformly drawn over* $\{0, 1\}$. *Performing counterfactual intervention ctf*(π) *leads to a submodel* $\mathcal{M}^*_\pi$ *described as a tuple*

$$\mathcal{M}^*_\pi = \left\langle \boldsymbol{U} = \{U_{i,1}, \dots, U_{i,4}\}, \boldsymbol{V}_\pi = \{X'_i, Y_i, S_i, X_i\}, \mathscr{F}_\pi, P(\boldsymbol{U})\right\rangle_{i=1,2,\dots}, \tag{373}$$

57. The connection between the counterfactual intervention and semi-autonomous systems was first formalized and explored in (Zhang and Bareinboim, 2022).

S_i	X_i	X'_i	$\mathbb{E}$	S_i	X_i	X'_i	$\mathbb{E}$
0	0	0	0.1	1	0	0	0.1
0	0	1	0.9	1	0	1	0.9
0	1	0	0.9	1	1	0	0.9
0	1	1	0.1	1	1	1	0.1

Table 20: Evaluation of the counterfactual expected reward $\mathbb{E}[Y_{i_{x_i}} \mid S_i, X_i = x'_i]$ evaluated in the MDP environment of Example 2.

The structural functions $\mathscr{F}_\pi$ are defined as, for every $i = 1, 2, \ldots,$

$$\mathscr{F}_i = \begin{cases} S_i \leftarrow \left(S_{i-1} \vee X'_{i-1}\right) \oplus U_{i-1,1} \oplus U_{i-1,2}, \\ X'_i \leftarrow S_i \oplus U_{i,1} \\ X_i \leftarrow S_i \oplus X'_i \oplus U_{i,4} \\ Y_i \leftarrow S_i \oplus X_i \oplus U_{i,1} \oplus U_{i,3} \end{cases} \tag{374}$$

Note that the structural function $X_i \leftarrow S_i \oplus U_{i,1}$. Given values of observed state $S_i = s_i$ and action $X_i = x_i$ in submodel $\mathcal{M}^*_\pi$, one could infer values of the unobserved confounder $U_{i,1}$ as

$$U_{i,1} = x_i \oplus s_i \tag{375}$$

Since the next state $S_{i+1} \leftarrow (S_i \vee X_i) \oplus U_{i,1} \oplus U_{i,3}$, given the current state $S_i = s_i$, intended action $X_i = x_i$, and realized action $X'_i = x'_i$, event $S_{i+1} = s_{i+1}$ implies the following

$$\begin{aligned} U_{i,3} &= s_{i+1} \oplus (s_i \vee x'_i) \oplus U_{i,1} && (376) \\ &= s_{i+1} \oplus (s_i \vee x'_i) \oplus x_i \oplus s_i && (377) \end{aligned}$$

The last step follows from Eq. 375. Evaluating the transition distribution on the next state S_{i+1} and next intended action X_{i+1} given the current state S_i, intended action X_i, and realized action X'_i in submodel $\mathcal{M}^*_\pi$ gives

$$\begin{aligned} &P\left(S_{i+1} = s_{i+1}, X_{i+1} = x_{i+1} \mid S_i = s_i, X_i = x_i, X'_i = x'_i\right) && (378) \\ &= P\left(U_{i,3} = s_{i+1} \oplus (s_i \vee x'_i) \oplus x_i \oplus s_i, X_{i+1} = x_{i+1} \mid S_i = s_i, X_i = x_i, X'_i = x'_i\right) && (379) \\ &= P\left(U_{i,3} = s_{i+1} \oplus (s_i \vee x'_i) \oplus x_i \oplus s_i, U_{i+1,1} = x_{i+1} \oplus s_{i+1}\right) && (380) \end{aligned}$$

The first step follows from Eq. 377; the second step follows from the equation $X'_{i+1} \leftarrow S_{i+1} \oplus U_{i+1,1}$. Moreover, given values of the current S_i, X'_i, X_i, the past history $S_1, \ldots, S_{i-1}$, $X'_1, \ldots, X'_{i-1}$, and $X_1, \ldots, X_{i-1}$ are independent from the exogenous variables $U_{i,3}, U_{i+1,1}$, i.e., the Markov property holds. We compute the detailed parametrization of the conditional transition distribution $P\left(S_{i+1}, X'_{i+1} \mid S_i, X'_i, X_i\right)$ and provide them in Table 19.

Similarly, note that values of the reward signal $Y_i \leftarrow S_i \oplus X_i \oplus U_{i,1} \oplus U_{i,2}$. Given state $S_i = s_i$, intended action $X'_i = x'_i$ and realized action $X_i = x_i$, event $Y_i = y_i$ implies the following

$$\begin{aligned} U_{i,2} &= y_i \oplus s_i \oplus x_i \oplus U_{i,1} && (381) \\ &= y_i \oplus s_i \oplus x_i \oplus s_i \oplus x'_i && (382) \\ &= y_i \oplus x_i \oplus x'_i && (383) \end{aligned}$$

The second step follows from Eq. 375. Evaluating the expected reward Y_i conditioning on the state S_i, intended action X'_i and realized action X_i in submodel $\mathcal{M}^*_\pi$ gives

$$\mathbb{E}\left[Y_i \mid S_i = s_i, X'_i = x'_i, X_i = x_i\right] \tag{384}$$
$$= P\left(Y_i = 1 \oplus x_i \oplus x'_i \mid S_i = s_i, X'_i = x'_i, X_i = x_i\right) \tag{385}$$
$$= P\left(U_{i,2} = 1 \oplus x_i \oplus x'_i\right) \tag{386}$$

Detailed parametrization of $\mathbb{E}\left[Y_i \mid s_i, x'_i, x_i\right]$ are computed and provided in Table 20. ■

More importantly, it is possible to show that the above conditional distributions in submodel coincide with ETTs of action X_i on the reward signal Y_i and next state S_{i+1} and action X_{i+1}, provided with the current state S_i. Such ETTs remain invariant across every stage $i = 1, \ldots, H$,

Lemma 2 *Let $\mathcal{M}$ be an MDP environment, $\pi(X_i \mid S_i, X'_i)$ be a counterfactual policy over actions $X_1, \ldots, X_H$, and $\mathcal{M}_\pi$ be an induced counterfactual submodel of $\mathcal{M}$. Then, for every $i = 1, \ldots, H$, the transition distribution over S_{i+1} and the expected reward over Y_i conditioning on S_i, X'_i, X_i in submodel $\mathcal{M}_\pi$ is equal to*

$$P\left(S_{i+1}, X_{i+1} \mid s_i, x'_i, x_i; \mathcal{M}_\pi\right) = P\left(S_{i+1_{X_i \leftarrow x_i}}, X_{i+1_{X_i \leftarrow x_i}} \mid s_i, X'_i = x'_i; \mathcal{M}\right) \tag{387}$$
$$\mathbb{E}\left[Y_i \mid s_i, x'_i, x_i; \mathcal{M}_\pi\right] = \mathbb{E}\left[Y_{i_{X_i \leftarrow x_i}} \mid s_i, X'_i = x'_i; \mathcal{M}\right] \tag{388}$$

Moreover, the above quantities remain invariant across stage $i = 1, \ldots, H$, i.e.,

$$P\left(S_{i+1_{X_i \leftarrow x_i}}, X_{i+1_{X_i \leftarrow x_i}} \mid s_i, X'_i = x'_i\right) = \cdots = P\left(S_{2_{X_1 \leftarrow x_1}}, X_{2_{X_1 \leftarrow x_1}} \mid s_1, X'_1 = x'_1\right) \tag{389}$$
$$\mathbb{E}\left[Y_{i_{X_i \leftarrow x_i}} \mid s_i, X'_i = x'_i\right] = \cdots = \mathbb{E}\left[Y_{1_{X_1 \leftarrow x_1}} \mid s_1, X'_1 = x'_1\right] \tag{390}$$

■

Among the above equations, Eqs. 387 and 388 follow from the definition of counterfactual intervention. Eqs. 389 and 390 hold since structural functions f_X, f_Y, f_S and the exogenous distribution $P(U_i)$ remain invariant across all stages of interventions $i = 1, \ldots, H$.

Example 55 *Consider the MDP environment $\mathcal{M}^*$ described in Eq. 5. Given values of state $S_i = s_i, X'_i = x'_i$, one could infer values of the unobserved confounder $U_{i,1}$ as*

$$U_{i,1} = x'_i \oplus s_i \tag{391}$$

Given current state and action $S_i = s_i, X'_i = x'_i$, the counterfactual event $S_{i+1_{X_i \leftarrow x_i}} = s_{i+1}$ implies

$$U_{i,3} = s_{i+1} \oplus (S_i \vee X'_i) \oplus U_{i,1} \tag{392}$$
$$= s_{i+1} \oplus (s_i \vee x_i) \oplus x'_i \oplus s_i \tag{393}$$

The last step follows from Eq. 391. Evaluating the ETT of action X_i on the next state S_{i+1} and action X_{i+1} conditioning on the current state S_i in the underlying MDP environment $\mathcal{M}^*$ gives

$$P\left(S_{i+1_{X_i \leftarrow x_i}} = s_{i+1}, X_{i+1_{X_i \leftarrow x_i}} = x_{i+1} \mid S_i = s_i, X'_i = x'_i\right) \tag{394}$$
$$= P\left(U_{i,3} = s_{i+1} \oplus (s_i \vee x_i) \oplus x'_i \oplus s_i, X_{i+1_{X_i \leftarrow x_i}} = x'_{i+1} \mid S_i = s_i, X'_i = x'_i\right) \tag{395}$$
$$= P\left(U_{i,3} = s_{i+1} \oplus (s_i \vee x_i) \oplus x'_i \oplus s_i, U_{i+1,1} = x'_{i+1} \oplus s_{i+1}\right) \tag{396}$$

The above equation coincides with Eq. 380. This means that the counterfactual ETT distribution $P\left(S_{i+1_{X_i \leftarrow x_i}}, X_{i+1_{X_i \leftarrow x_i}} \mid s_i, x_i'\right)$ *evaluated in the MDP environment* $\mathcal{M}^*$ *is equal to the conditional transition distribution* $P\left(S_{i+1}, X_{i+1}' \mid s_i, x_i', x_i\right)$ *evaluated in submodel* $\mathcal{M}_\pi^*$*. Its detailed parametrizations are provided in Table 19. Moreover, this counterfactual distribution remains the same for all decision horizon* $i = 1, 2, \ldots$ *since structural functions* $\mathscr{F}$ *and exogenous distribution* $P(U_{i,1}, U_{i,2}, U_{i,3})$ *are invariant with regard to the horizon* i*.*

Similarly, given state and action $S_i = s_i, X_i' = x_i'$*, the potential outcome* $Y_{i_{X_i \leftarrow x_i}} = y_i$ *implies*

$$U_{i,2} = y_i \oplus s_i \oplus x_i \oplus U_{i,1} \tag{397}$$
$$= y_i \oplus s_i \oplus x_i \oplus s_i \oplus x_i' \tag{398}$$
$$= y_i \oplus x_i \oplus x_i' \tag{399}$$

The second step follows from Eq. 391. Evaluating the ETT of action X_i *on the reward signal* Y_i *conditioning on the current state* S_i *in the underlying MDP environment* $\mathcal{M}^*$ *gives*

$$\mathbb{E}\left[Y_{i_{X_i \leftarrow x_i}} \mid S_i = s_i, X_i' = x_i'\right] = P\left(Y_{i_{X_i \leftarrow x_i}} = 1 \oplus x_i \oplus x_i' \mid S_i = s_i, X_i' = x_i'\right) \tag{400}$$
$$= P\left(U_{i,2} = 1 \oplus x_i \oplus x_i'\right) \tag{401}$$

The last step coincides with Eq. 386. This means that the conditional x*-specific causal effects* $\mathbb{E}\left[Y_{i_{X_i \leftarrow x_i}} \mid s_i, x_i'\right]$ *evaluated in the MDP environment* $\mathcal{M}^*$ *equates to the conditional expected reward* $\mathbb{E}[Y_i \mid s_i, x_i', x_i]$ *evaluated in submodel* $\mathcal{M}_\pi^*$*. Its detailed parametrizations remain invariant for every decision horizon* $i = 1, 2, \ldots$*, and are provided in Table 20.* ■

Lem. 2 permits us to represent the distribution $P_\pi\left(\boldsymbol{S}, \boldsymbol{X}', \boldsymbol{X}, \boldsymbol{Y}\right)$ induced by a counterfactual policy $\pi = (\pi_1(X_1 \mid S_1, X_1'), \ldots, \pi_H(X_H \mid S_H, X_H'))$ using a standard MDP (Def. 12)

$$\langle \mathscr{D}(S) \times \mathscr{D}(X), \mathscr{D}(X), \mathcal{T}_{\text{ctf}}, \mathcal{R}_{\text{ctf}} \rangle \tag{402}$$

Here, $\mathscr{D}(S)$ and $\mathscr{D}(X)$ are, respectively, the domain of state S_i and action X_i for every stage $i = 1, \ldots, H$. The transitional distribution $\mathcal{T}_{\text{ctf}}$ and the reward function $\mathcal{R}_{\text{ctf}}$ are conditional ETTs evaluated in the underlying MDP environment $\mathcal{M}^*$ given by, for any $s, s' \in \mathscr{D}(S)$ and any $x, x', x'' \in \mathscr{D}(X)$,

$$\mathcal{T}_{\text{ctf}}(s, x, x', s', x'') = P\left(S_{i+1_{X_i \leftarrow x'}} = s', X_{i+1_{X_i \leftarrow x'}} = x'' \mid S_i = s, X_i = x\right) \tag{403}$$
$$\mathcal{R}_{\text{ctf}}(s, x, x') = \mathbb{E}\left[Y_{i_{X_i \leftarrow x'}} \mid S_i = s, X_i = x\right] \tag{404}$$

Let $\mathcal{R}(\boldsymbol{Y}) \in \mathbb{R}$ be a reward function taking reward signal $\boldsymbol{Y} = \{Y_1, \ldots, Y_H\}$ as input. Our goal is to obtain an optimal counterfactual policy $\pi^* \in \Pi_{\text{CTF}}$ maximizing the expected reward over $\mathbb{E}_\pi\left[\mathcal{R}(\boldsymbol{Y})\right]$ evaluated in the underlying MDP environment $\mathcal{M}^*$. The reduction to a standard MDP model $\langle \mathscr{D}(S) \times \mathscr{D}(X), \mathscr{D}(X), \mathcal{T}_{\text{ctf}}, \mathcal{R}_{\text{ctf}} \rangle$ allows us to solve for an optimal counterfactual policy π^* using standard dynamic programming algorithms (Bellman, 1966), provided that the detailed parametrization of the underlying MDP environment $\mathcal{M}^*$ is available.

To make this argument more precise, we will consider a discounted cumulative reward function $\mathcal{R}(\boldsymbol{Y}) = \sum_{i=1}^{\infty} \gamma^{i-1} Y_i$ over an infinite horizon $H = \infty$, where the discount rate $\gamma \in (0, 1)$. The

S_i	X_i	X_i'	Q_*	S_i	X_i	X_i'	Q_*
0	0	0	1	1	0	0	1
0	0	1	**9**	1	0	1	**9**
0	1	0	**9**	1	1	0	**9**
0	1	1	1	1	1	1	1

Table 21: Optimal augmented Q-function $Q_*(s, x, x')$ evaluated in the MDP model of Example 2.

classic result in planning literature (Puterman, 1994) implies that it is sufficient to consider a class of stationary counterfactual policies $\pi = (\pi_1(X_1 \mid S_1, X_1), \pi_2(X_2 \mid S_2, X_2), \dots)$ such that the decision rule π_i remains invariant across the decision horizon $i = 1, 2, \dots$, i.e., $\pi_1 = \pi_2 = \dots$.

The state-action value function $Q_\pi : \mathscr{D}(S) \times \mathscr{D}(X) \times \mathscr{D}(X) \to \mathbb{R}$ for a counterfactual policy π evaluated in the MDP environment $\mathcal{M}^*$ is defined as the expected cumulative reward following policy π, given the starting state s, intended action x, and realized action x'. Formally,

$$Q_\pi(s, x, x') = \mathbb{E}\left[\sum_{j=0}^{\infty} \gamma^j Y_{i+j} \mid S_i = s, X_i = x, X_i' = x'; \mathcal{M}_\pi^*\right] \tag{405}$$

By exploring the Markov property in submodel $\mathcal{M}_\pi^*$ (Eqs. 370 and 371, the above Q-function could be further written as an augmented *Bellman Equation* using the intended action X as an additional side information:

$$\begin{aligned}
&Q_\pi(s, x, x') && (406)\\
&= \mathbb{E}\left[Y_i + \gamma Y_{i+1} + \gamma^2 Y_{i+2} + \cdots \mid S_i = s, X_i = x, X_i' = x'; \mathcal{M}_\pi^*\right] && (407)\\
&= \mathbb{E}\left[Y_i + \gamma Q_\pi(S_{i+1}, X_{i+1}, X_{i+1}') \mid S_i = s, X_i = x, X_i' = x'; \mathcal{M}_\pi^*\right] && (408)\\
&= \mathcal{R}_{\text{ctf}}(s, x, x') + \gamma \sum_{s', x''} \mathcal{T}_{\text{ctf}}(s, x, x', s', x'') \sum_{x'''} \pi_{i+1}(x''' \mid s', x'') Q_\pi(s', x'', x''') && (409)
\end{aligned}$$

The last step follows from the equality relationships in Lem. 2. An optimal counterfactual policy π^* is obtainable by recursively computing the Q-function and optimizing the realized action x' for every state s and intended action x. The following example demonstrates such a procedure.

Example 56 *Consider the MDP environment $\mathcal{M}^*$ defined in Eq. 5. Note that an optimal counterfactual policy π^* is such that its induced value function $Q_{\pi^*}(s, x, x') \geq Q_\pi(s, x, x')$ for all state s, intended action x, and realized action x'. Optimizing Q-function leads to a counterfactual augmented (ctf-augmented)* Bellman optimality equation *using the intended action X as an additional context, i.e,*

$$Q_*(s, x, x') = \mathcal{R}_{ctf}(s, x, x') + \gamma \sum_{s', x''} \mathcal{T}_{ctf}(s, x, x', s', x'') \max_{x'''} Q_*(s', x'', x''') \tag{410}$$

The optimal policy π^* is given by, for every stage $i = 1, 2, \dots$, for any state $s \in \mathscr{D}(S)$,

$$\pi^*(S_i = s, X_i = x) = \arg\max_{x'} Q_*(s, x, x') \tag{411}$$

We compute the optimal augmented Q-function evaluated in the MDP environment $\mathcal{M}^*$ using the value iteration algorithm (Sutton and Barto, 1998). Detailed parametrizations are provided in Table 21. The optimal policy $\pi^* = (\pi_i^*(X_i \mid S_i, X_i))_{i=1}^{\infty}$ is given by $\pi_i^* \triangleq X_i \leftarrow \neg X_i$, for every $i = 1, 2, \ldots$.

After all, this means that, in this case, the agent should always go against its intended action. Evaluating its expected return gives $\mathbb{E}_{\pi^*}\left[\sum_{i=1}^{\infty} \gamma^{i-1} Y_i\right] = 9$, which outperforms the best possible experimental policy $\pi_i \triangleq X_i \leftarrow \neg S_i$. Derivations of the best experimental policy are provided in Example 16. ■

More generally, experimental policies in $\Pi_{\text{EXP}} = \{\langle X_i, \{S_i\}\rangle\}_{i=1}^{H}$ are contained in the counterfactual policy space $\Pi_{\text{CTF}} = \{\langle X_i, \{S_i, X_i\}\rangle\}_{i=1}^{H}$. This means that one could simulate the expected reward of any experimental policy $\pi' = (\pi'_i(X_i \mid S_i))_{i=1}^{H}$ using a counterfactual policy $\pi = (\pi_i(X_i \mid S_i, X_i))_{i=1}^{H}$ such that $\pi'_i(x' \mid s) = \pi_i(x' \mid s, x)$ for all s, x, x'. In this case, intended actions $\boldsymbol{X}$ do not affect values of reward signals $\boldsymbol{Y}$ in submodel $\mathcal{M}_{\pi}^*$ induced by counterfactual intervention $\text{ctf}(\pi)$. Marginalizing intended actions $\boldsymbol{X}$ reduces $\mathcal{M}_{\pi}$ to the experimental submodel $\mathcal{M}_{\pi'}^*$ induced by intervention $\text{do}(\pi')$. The performance of the counterfactual policy π and the experimental policy π' thus coincides, i.e., $\mathbb{E}_{\pi}[\mathcal{R}(\boldsymbol{Y})] = \mathbb{E}_{\pi'}[\mathcal{R}(\boldsymbol{Y})]$. This observation implies that an agent optimizing the environment following counterfactual interventions $\text{ctf}(\pi)$ must perform at least as well as its counterpart following experimental intervention $\text{do}(\pi)$. Formally,

Theorem 16 (Counterfactual dominates Interventional policies) *For an MDP environment $\mathcal{M}^*$, let $\mathcal{R}$ be a reward function over reward signals $\boldsymbol{Y} = \{Y_1, \ldots, Y_H\}$. Let policy spaces $\Pi_{\text{CTF}} = \{\langle X_i, \{S_i, X_i\}\rangle\}_{i=1}^{H}$ and $\Pi_{\text{EXP}} = \{\langle X, \{S_i\}\rangle\}_{i=1}^{H}$. Then,*

$$\max_{\pi \in \Pi_{\text{CTF}}} \mathbb{E}_{\pi}[\mathcal{R}(\boldsymbol{Y})] \geq \max_{\pi \in \Pi_{\text{EXP}}} \mathbb{E}_{\pi}[\mathcal{R}(\boldsymbol{Y})] \tag{412}$$

■

Thm. 16 implies that it is preferable to learn an optimal counterfactual policy in the MDP environment when determining values for a sequence of actions, which consistently dominates the best possible experimental policy that does not account for the agent's intended actions X_i for every stage of decision $i = 1, \ldots, H$. On the other hand, when the NUC holds in $\mathcal{M}^*$ (Def. 13), conditioning on current state S_i and realized action X_i d-separates the intended action X_i from all the future rewards $Y_i, Y_{i+1}, \ldots$ and states $S_{i+1}, S_{i+2}, \ldots$. We thus have the following, for any x_i, s_i, x'_i,

$$P\left(S_{i+1_{X_i \leftarrow x'_i}}, X_{i+1_{X_i \leftarrow x'_i}} \mid S_i = s_i, X_i = x_i\right) = P_{X_i \leftarrow x'_i}(S_{i+1} \mid S_i = s_i) \tag{413}$$

$$\mathbb{E}\left[Y_{i_{X_i \leftarrow x'_i}} \mid s_i, X_i = x_i\right] = \mathbb{E}_{X_i \leftarrow x'_i}[Y_i \mid S_i = s_i] \tag{414}$$

In words, the counterfactual transition probabilities $\mathcal{T}_{\text{ctf}}$ and reward function $\mathcal{R}_{\text{ctf}}$ coincide with their experimental counterparts $\mathcal{T}_{\text{exp}}$ and $\mathcal{R}_{\text{exp}}$. An agent can thus simulate the performance of an optimal counterfactual policy using an experimental one; observing the agent's intended action provides no value of information to the learning task and could be ignored.

7.2 Counterfactual Randomization

So far, we have described effective planning algorithms to obtain an optimal counterfactual policies that account for the agent's intended actions in canonical decision environments, such as MABs

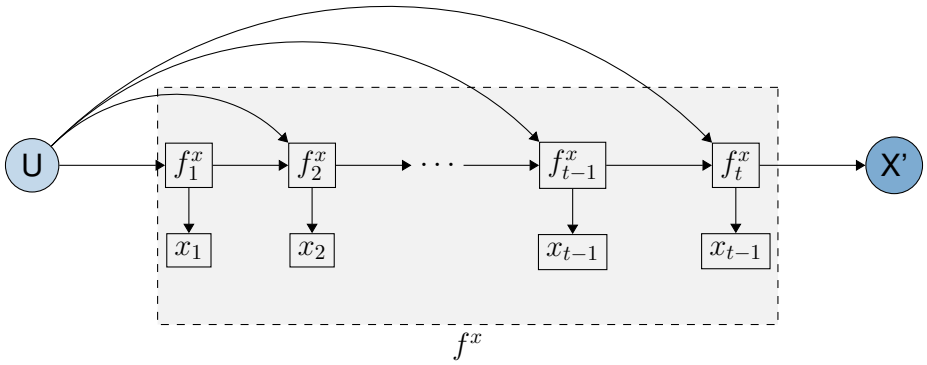


Figure 45: Illustration of decision flow, f_X, where U is taken as input and the natural predilections X' is returned as output. The process is refined through multiple stages.

and MDPs. However, how can an optimal counterfactual policy be learned by computing the counterfactual quantities entailed by the underlying, unknown environment? To illustrate, consider the MAB environment as an example. When the intended action (coming from f_x) and the executed action match (i.e., $x' = x$), the counterfactual quantity $\mathbb{E}[Y_x \mid x]$ coincides with the observational reward $\mathbb{E}[Y \mid X = x]$, following the composition axiom (Pearl, 2000, Ch. 7.3). For the general case when intended and executed actions do not match ($x' \neq x$), the x-specific effect $\mathbb{E}[Y_x \mid x']$ is not computable from any combination of passive observations and controlled experiments, without the detailed parametrization or additional assumptions of the underlying causal model.[58]

In settings where the x-specific effect is identifiable, a more challenging question arises: How can an agent implement the counterfactual policy in the environment? This question stems from the observation that agents may consider various alternatives during the deliberation process and change their opinion about the best course of action. Consequently, only the final choice matters, actually representing the agent's natural predilections. To illustrate this concept, the diagram in Fig. 45 depicts an example of an agent's deliberation process. Initially, the agent intends to play $X' = x_1$ but reconsiders, thinking it might be sub-optimal, and decides to switch to $X' = x_2$, where $x_1 \neq x_2$. As time passes, the agent may realize that $X' = x_{t-1}$ was not ideal and switch to an alternative, $X' = x_t$. Ultimately, the final decision defines the agent's individuality, irrespective of the path taken to reach it.[59]

This challenge calls for novel counterfactual machinery to allow for the counterfactual interaction following layer 3 as discussed in the previous section, in theory. Here, we introduce a novel type of randomization for intention-specific groups, namely, interrupt any reasoning agent before they execute their choice, treat this choice as their intention, deliberate, and then act. The x-specific effect will then be computed in an alternative fashion, based on the idea of intention-specific randomization. This section discusses the algorithmic implementation of this randomization.

Alg. 8 shows the detailed design of randomized controlled trials augmented with counterfactual interventions, which we name `Ctf-RCT`. More specifically, it takes as input the domain of action X

58. One exception is the binary case, as elaborated in (Pearl, 2000, Sec. 8.2).

59. Note that whenever the agent pursues an interventional (layer 2) strategy by leveraging Fisherian randomization, the entire deliberation process is bypassed. In this case, the randomization itself determines which action should be executed. This is illustrated in Fig. 45.

Algorithm 8 Counterfactual Randomized Controlled Trials (`Ctf-RCT`) in MAB

Require: the domain of action $\mathscr{D}(X)$, the total number of trials $N \in \mathbb{N}$.

1: **for** episodes $t = 1, 2, \ldots$ **do**

2: Perceive an intended action $X^{(t)}$ and store it.

3: Choose a realize action $X'^{(t)}$ as follows.

$$X'^{(t)} = \begin{cases} \texttt{Unif}(\mathscr{D}(X)) & \text{if } t \leq N \\ \arg\max_x \hat{\mathbb{E}}^{(N)} \left[Y_{X \leftarrow x} \mid X = X^{(t)} \right] & \text{if } t > N \end{cases} \tag{416}$$

4: Perform $\text{do}(X'^{(t)})$ for episode t and receive reward $Y^{(t)}$.

5: **end for**

and an integer N indicating the total number of trials. For every episode t, the agent first perceives its intended action $X^{(t)}$, decided by the behavioral policy f_X. During the exploration phase (i.e., episode $t \leq N$), the algorithm selects a realized action $X^{(t)}$ from the action space $\mathcal{D}(X)$ uniformly at random. During the exploitation phase (episode $t > N$), it selects a realized action maximizing the empirical estimates of the x-specific effect provided with the intended action $X = X^{(t)}$, computed from samples collected from the first N episodes of interventions. Formally, the empirical estimates of the counterfactual quantity $\mathbb{E}\left[Y_{X \leftarrow x} \mid X = x'\right]$ computed from samples up to episode t are defined as:

$$\hat{\mathbb{E}}^{(t)} \left[Y_{X \leftarrow x} \mid X = x' \right] = \frac{\sum_{i=1}^{t} Y^{(i)} \mathbb{1} \left\{ X'^{(i)} = x', X^{(i)} = x \right\}}{N_t\left(x', x\right)} \tag{415}$$

where $N_t(x, x') = \sum_{i=1}^{t} \mathbb{1} \left\{ X^{(i)} = x, X'^{(i)} = x \right\}$ is the total number of occurrence of intercepting intended actions $X^{(i)} = x$ and selecting realized actions $X'^{(i)} = x'$. It follows that Eq. 415 is a consistent estimate of the x-specific effect of X on Y. Finally, `Ctf-RCT` performs an intervention $\text{do}(X'^{(t)})$ following the selected action throughout episode t and receives subsequent reward $Y^{(t)}$.

One could also apply the same principle of counterfactual interventions to adaptive online randomization algorithms such as `UCB` (Alg. 3) to obtain an optimal counterfactual policy with sublinear regret. Precisely, by applying Hoeffiding's inequalities for every intended action $X = x'$, we define the upper confidence bound over ETT $\mathbb{E}\left[Y_{X \leftarrow x} \mid X = x'\right]$ computed from samples collected up to episode t as follows, for every pair $x', x \in \mathscr{D}(X)$,

$$\text{UCB}_t\left(x, x', \delta\right) = \hat{\mathbb{E}}^{(t)}\left[Y_{X \leftarrow x'} \mid X = x\right] + \sqrt{\frac{\log(1/\delta)}{2N_t(x, x')}}, \tag{417}$$

where the error probability $\delta \in (0, 1)$ is an arbitrary real value. For every episode $t = 1, 2, \ldots$, the algorithm incorporating counterfactual interventions first perceives and intercepts the agent's intended action $X^{(t)}$. It then computes the confidence bounds $\text{UCB}_{t-1}\left(x, X^{(t)}, \delta\right)$ for every arm x' from samples collected up to episode $t-1$, and picks a realized action $X^{(t)}$ with the highest UCB estimates provided with the intended action $X'^{(t)}$. Finally, the algorithm performs an intervention $\text{do}(X^{(t)})$ following the selected action throughout episode t and receives subsequent reward $Y^{(t)}$.

The detailed implementation of the counterfactual `UCB` algorithm, named `Ctf-UCB`, is summarized in Alg. 9. For every episode $t = 1, 2, \ldots$, the error probability δ is set as a non-increasing

Algorithm 9 Counterfactual Upper Confidence Bound (`Ctf-UCB`) in MAB

Require: the domain of action $\mathscr{D}(X)$.

1: **for** episodes $t = 1, 2, \ldots$ **do**
2: Receive an intended action $X^{(t)}$.
3: Choose an arm $X'^{(t)} = \arg\max_x \text{UCB}_{t-1}\left(x, X^{(t)}, \delta\right)$ where $\delta = N_t\left(X^{(t)}\right)^{-4}$.
4: Perform $\text{do}(X'^{(t)})$ for episode t and receive reward $Y^{(t)}$.
5: **end for**

function of the total occurrences $N_t\left(X^{(t)}\right) = \sum_{j=1}^{t} \mathbb{1}\left\{X^{(j)} = X^{(t)}\right\}$ of the agent's intended action perceived at episode t from past samples collected so far. In words, the algorithm uses a separate instance of `UCB` for every intended action $X = x$. Let $K = |\mathscr{D}(X)|$ denote the total number of candidate arms. The cumulative regret of `Ctf-UCB` after T episodes of interventions is bounded by summing the regrets of each `UCB` instance for every intended action $X = x$. Specifically, summing regrets in Eq. 185 gives

$$R(T, \mathcal{M}^*) \leq \sum_x C \sqrt{K \sum_{t=1}^{T} \mathbb{1}\left\{X^{(t)} = x\right\} \log(T)}, \tag{418}$$

where C is a universal constant. Since the square root is a concave function, applying Jensen's inequality allows us to further bound the regret as follows.

Theorem 17 (Regrets of `Ctf-UCB` in MABs) *For an MAB $\langle \mathcal{M}^*, \Pi, Y \rangle$, let Π be a counterfactual policy space $\{\langle X, \{X\}\rangle\}$, Y be the reward variable with support on $[0, 1]$, and let the domain of action X be $\mathscr{D}(X) = \{1, \ldots, K\}$. The regret of `Ctf-UCB` in SCM $\mathcal{M}^*$ after $T > 1$ episodes is bounded by*

$$R(T, \mathcal{M}^*) \leq CK\sqrt{T \log(T)} \tag{419}$$

where C is a universal constant. ■

Thm. 17 implies that `Ctf-UCB` is able to eventually learn an optimal counterfactual policy $\pi^*(X \mid X')$ that consistently improves the agent's intended action. On the other hand, without perceiving the agent's intended action, the standard `UCB` algorithm only performs atomic interventions $\text{do}(x)$. Note that an optimal counterfactual policy consistently dominates the best possible interventional one (Thm. 15). One important observation is that the standard `UCB` generally experiences linear regret when compared with an optimal counterfactual agent.

Corollary 4 *Let Π be an experimental policy space $\{\langle X, \emptyset\rangle\}$, Y be the reward variable with support on $[0, 1]$, and let the domain of action X be $\mathscr{D}(X) = \{1, \ldots, K\}$. There exists an MAB environment $\mathcal{M}^*$ such that for any algorithm (e.g., `UCB`) optimizing over space Π after $T > 1$ episodes is lower bounded by*

$$R(T, \mathcal{M}^*) \geq 0.5T \tag{420}$$

■

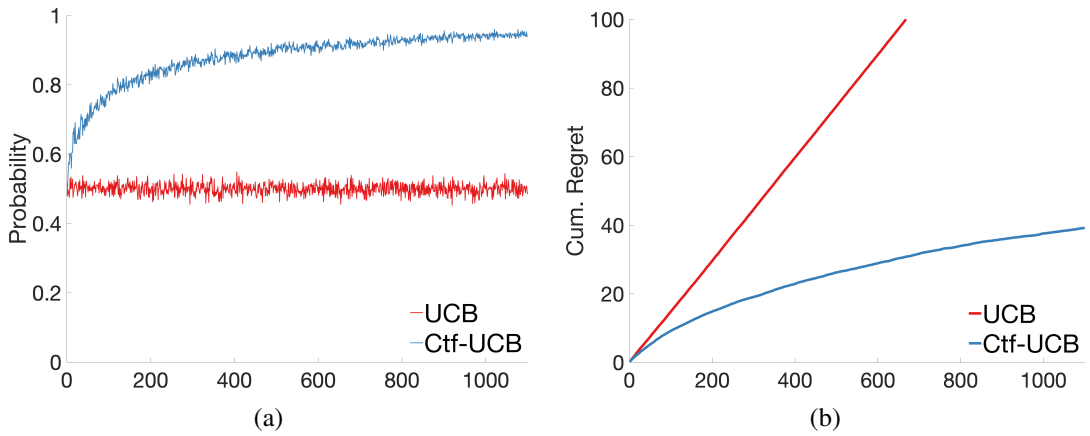


Figure 46: Performance of standard `UCB` performing atomic interventions and the augmented `Ctf-UCB` using counterfactual interventions; x-axis represents the total episodes of interactions. The x-axis represents, respectively, the probability of picking an optimal action and the cumulative regret in (a) and (b); the y-axis represents the number of episodes in both (a) and (b).

In words, there is an MAB environment such that for any online algorithm employing Fisherian randomization, it must incur at least 0.5 regret on average per episode of interaction. It is thus unable to achieve an optimal counterfactual policy accounting for the agent's intuition.[60]

The proposed augmentation procedure may appear to be an immediate extension of `UCB` in contextual bandits using the agent's natural predilection as an extra context. However, the augmented `Ctf-UCB` differs from contextual `UCB` in the following.[61]

1. The agent's intended action X is semantically different from a context S. The former is a variable only existing under the observational regime (see), as shown in Fig. 43a. It is replaced by the agent's realized action and does not appear under the interventional regime (do), as shown in Fig. 43b. On the other hand, the context variable S is not affected by the agent's interaction regime with the environment.

2. The realization of using the agent's intended action X as an additional context is a consequence of counterfactual intervention (ctf-do). On the other hand, online RL algorithms interacts with the environment by repeatedly performing randomized interventions (do), discarding the agent's natural predilection.

Experiment 9 *We evaluate the standard `UCB` algorithm that attempts to maximize rewards based on $\mathbb{E}_x[Y]$, ignoring the agent's intended arm choice, and `Ctf-UCB` described in Alg. 9, which maximizes the rewards based on ETT $\mathbb{E}[Y_{X\leftarrow x'} \mid X = x]$ via counterfactual interventions. All reported simulations are partitioned into rounds of $T = 1,000$ trials averaged over $N = 1,000$ repetitions.*

The Greedy Casino parameterization (specified in Table 16) illustrates the scenario where each arm's payout appears to be equivalent under the observational and experimental distributions

60. See Example 51 for details of the construction of this MAB environment.

61. This augmentation procedure is applicable to empower other bandit algorithms, including Thompson sampling (Bareinboim et al., 2015; Forney et al., 2017; Forney and Bareinboim, 2019), with the capability of counterfactual randomization and obtain an optimal counterfactual policy.

alone. Only when we concert the two distributions and condition on a player's predilection the optimal policy can be obtained. Fig. 46 shows the cumulative regret and probability of selecting optimal arms for evaluated algorithms. Simulations support the efficacy of counterfactual interventions. Analyses revealed a significant difference in the regret experienced by `Ctf-UCB` *compared to standard* `UCB`*, which, predictably, is not a competitor experiencing linear regret.* ■

So far we have introduced a novel type of interaction between the agent and the environment, i.e., the counterfactual randomization, in single-stage decision-making settings such as MABs. We showed online learning agents using counterfactual randomizations consistently outperform their experimental counterparts that do not actively consider the agent's intended action. Novel adaptive counterfactual randomization procedures were proposed to optimize an unknown MAB environment.

7.2.1 COUNTERFACTUAL RANDOMIZATION FOR MDPS

In this section, we generalize the counterfactual randomization to the more generalized sequential decision-making setting, e.g., MDPs. We will endow online algorithms in an unknown MDP with the ability of counterfactual reasoning so that they can learn optimal counterfactual policies, accounting for the agent's intended action and natural predilections.

To make the argument more precise, we will focus on the episodic learning setting in an unknown MDP environment $\mathcal{M}^*$ with a finite horizon H. The algorithm will interact with $\mathcal{M}^*$ for repeated episodes $t = 1, 2, \ldots, T$. For every episode t, the algorithm picks a counterfactual policy $\pi^{(t)} = \left(\pi_1^{(t)}(X_1 \mid S_1, X_1'), \ldots, \pi_H^{(t)}(X_H \mid S_H, X_H')\right)$ in the counterfactual space Π_{CTF}, performs intervention ctf $\left(\pi^{(t)}\right)$, and receives subsequent reward signals $\boldsymbol{Y}^{(t)} = \left\{Y_1^{(t)}, \ldots, Y_H^{(t)}\right\}$. We are interested in maximizing the undiscounted cumulative reward $\mathcal{R}(\boldsymbol{Y}) = \sum_{i=1}^{H} Y_i$.[62] For the convenience of the analysis, we will assume that parameters of the counterfactual reward function $\mathcal{R}_{\text{ctf}}(s, x, x') = \mathbb{E}\left[Y_{i_{X_i \leftarrow x'}} \mid S_i = s, X_i = x\right]$ are known. However, our analysis generalizes immediately to settings where the reward function is not accessible.

We will utilize `UCBVI` (Azar et al., 2017), an online reinforcement learning algorithm that can learn the best possible experimental policy $\pi \in \Pi_{\text{EXP}}$ in a finite-horizon MDP environment. Alg. 10 shows an augmented procedure that incorporates counterfactual randomization which we call `Ctf-UCBVI`. More specifically, for every episode t, it computes a policy $\pi^{(t)}$ based on the data $\boldsymbol{H}^{(t-1)}$ collected prior to episode t. At Step 3, it calls `UCB-Q-values` (Alg. 11), which returns upper confidence bounds on the optimal Q-values $Q_*(s, x, x')$. This is computed using an empirical Bellman operator with an additional confidence bonus, estimated based on Chernoff-Hoeffding's concentration inequality. The empirical estimates of the counterfactual transition distributions $\mathcal{T}_{\text{ctf}}$ and the conditional reward $\mathcal{R}_{\text{ctf}}$ are consistent following Lem. 2. At Step 6 in `UCB-Q-values`, the linear operator $\left(\hat{\mathcal{T}}_t \cdot V_t^{(h+1)}\right)(s, x, x')$ is defined as,

$$\left(\hat{\mathcal{T}}_t \cdot V_t^{(h+1)}\right)(s, x, x') = \sum_{s', x''} \hat{\mathcal{T}}_t(s, x, x', s', x'') V_t^{(h+1)}(s', x'') \tag{426}$$

At Steps 5 – 9, `Ctf-UCBVI` sequentially performs counterfactual intervention ctf on every action $X_1, \ldots, X_H$. For every decision horizon $i = 1, \ldots, H$, it observes the current state $S_i = S_i^{(t)}$, and

62. If the decision horizon H is sufficiently large, the undiscounted cumulative reward provides an approximation to the discounted cumulative reward with an infinite horizon (Kearns et al., 1999).

Algorithm 10 Counterfactual `UCBVI` in MDP (`Ctf-UCBVI`)

Require: a policy space $\Pi = \{\langle X_i, \{S_i, X_i\}\rangle\}_{i=1}^{H}$, a reward function $\mathcal{R}(\boldsymbol{Y}) = \sum_{i=1}^{H} Y_i$.

Initialize data $\boldsymbol{H}^{(0)} = \emptyset$.
for episodes $t = 1, 2, \ldots$ **do**
  $\hat{Q}_{t-1} = \texttt{UCB-Q-values}(\boldsymbol{H}^{(t-1)})$.
  for step $i = 1, \ldots, H$ **do**
    Observe current state $S_i = S_i^{(t)}$.
    Intercept the agent's intended action $X_i = X_i^{(t)}$.
    Pick a new action $X_i'^{(t)} = \arg\max_x \hat{Q}_{t-1}^{(i)}\left(S_i^{(t)}, X_i^{(t)}, x\right)$.
    Perform do$(X_t \leftarrow X_i'^{(t)})$ and receive reward $Y_i^{(t)}$.
    Update data $\boldsymbol{H}^{(t)} = \boldsymbol{H}^{(t-1)} \cup \left\{S_i^{(t)}, X_i^{(t)}, X_i'^{(t)}\right\}$.
  end for
end for

intercepts the agent's intended action $X_i' = X_i^{(t)}$. The algorithm then computes an alternative action $X_i^{(t)}$ by maximizing the empirical Q-values $\hat{Q}_{t-1}$ computed from data $\boldsymbol{H}^{(t)}$. Finally, it performs the selected action do$(X_t \leftarrow X_i'^{(t)})$ and receives a subsequent reward $Y_i^{(t)}$.

Following the derivation in (Azar et al., 2017), it is possible to show that `Ctf-UCBVI`, empowered with counterfactual randomization, is able to obtain an optimal counterfactual policy in Π_{CTF} while achieving a sublinear regret. Formally,

Theorem 18 (Regrets of `Ctf-UCBVI` in MDPs) *For an MDP $\langle \mathcal{M}^*, \Pi, \mathcal{R}\rangle$ with horizon $H \in \mathbb{N}$, let Π be a counterfactual policy space $\{\langle X_i, \{S_i, X_i\}\rangle\}_{i=1}^{N}$, $\mathcal{R}(\boldsymbol{Y}) = \sum_{i=1}^{H} Y_i$ be a cumulative reward function over bounded reward signals $Y_i \in [0, 1]$. It holds the regret of `Ctf-UCBVI` in SCM $\mathcal{M}^*$ after $T > 1$ episodes is bounded by*

$$R(T, \mathcal{M}^*) \leq CH^{3/2}\sqrt{|\mathscr{D}(S) \times \mathscr{D}(X)|T\log(T)} \tag{427}$$

where C is a universal constant; $\mathscr{D}(S)$ and $\mathscr{D}(X)$ are domains of every state S_i and action X_i, $i = 1, 2, \ldots$, respectively. ■

On the other hand, without considering the agent's intended action, online algorithms performing randomized experiments may never be able to converge to an optimal counterfactual policy. The linear regret could occur when the intended action X_i reveals valuable information about the unobserved confounder U_i, as highlighted by the next proposition.

Corollary 5 *Let Π be an experimental policy space $\{\langle X_i, \{S_i, X_i\}\rangle\}_{i=1}^{H}$, $\mathcal{R}(\boldsymbol{Y}) = \sum_{i=1}^{H} Y_i$ be a reward function over bounded reward signal $Y_i \in [0, 1]$. There exists an MDP environment $\mathcal{M}^*$ such that for any algorithm (e.g., `UCBVI`) optimizing over space Π after $T > 1$ episodes is lower bounded by*

$$R(T, \mathcal{M}^*) \geq 0.08HT \tag{428}$$

■

Algorithm 11 `UCB-Q-values`

Require: Data $\boldsymbol{H}^{(t)} = \left\{\boldsymbol{S}^{(i)}, \boldsymbol{X}^{(i)}, \boldsymbol{X}'^{(i)}, \boldsymbol{Y}^{(i)}\right\}_{i=1}^{t}$

1: For all $s, s' \in \mathscr{D}(S)$ and all $x, x', x'' \in \mathscr{D}(X)$, compute from data $\boldsymbol{H}^{(t)}$

$$N_t(s, x, x') = \sum_{k=1}^{t}\sum_{i=1}^{H} \mathbb{1}\left\{S_i^{(k)} = s, X_i^{(k)} = x, X_i'^{(k)} = x'\right\} \tag{421}$$

$$N'_t(s, x, x') = \sum_{k=1}^{t}\sum_{i=1}^{H} Y_i^{(k)}\mathbb{1}\left\{S_i^{(k)} = s, X_i^{(k)} = x, X_i'^{(k)} = x'\right\} \tag{422}$$

$$N_t(s, x, x', s', x'') = \sum_{k=1}^{t}\sum_{i=1}^{H} \mathbb{1}\left\{S_i^{(k)} = s, X_i^{(k)} = x, X_i'^{(k)} = x', S_{i+1}^{(k)} = s, X_{i+1}'^{(k)} = x''\right\} \tag{423}$$

2: Let $\mathcal{K} = \{(s, x, x') \in \mathscr{D}(S) \times \mathscr{D}(X) \times \mathscr{D}(X) \mid N_t(s, x, x') > 0\}$.

3: For $(s, x, x') \in \mathcal{K}$, compute estimates

$$\hat{\mathcal{T}}_t(s, x, x', s', x'') = \frac{N_t(s, x, x', s', x'')}{N_t(s, x, x')}, \qquad \hat{\mathcal{R}}_t(s, x, x') = \frac{N'_t(s, x, x')}{N_t(s, x, x')} \tag{424}$$

4: Initialize $V_t^{(H+1)}(s, x) = 0$, $Q_t^{(H)}(s, x, x') = H$ for all $(s, x, x') \in \mathscr{D}(S) \times \mathscr{D}(X) \times \mathscr{D}(X)$.

5: **for** $i = H, H-1, \ldots, 1$ **do**

6: For all $(s, x, x') \in \mathcal{K}$, compute function $Q_t^{(h)}$ as

$$Q_t^{(h)}(s, x, x') = \min\left\{Q_{t-1}^{(h)}(s, x, x'), H, \hat{\mathcal{R}}_t(s, x, x') + \left(\hat{\mathcal{T}}_t V_t^{(h+1)}\right)(s, x, x') + b_t^{(h)}(s, x, x')\right\}$$

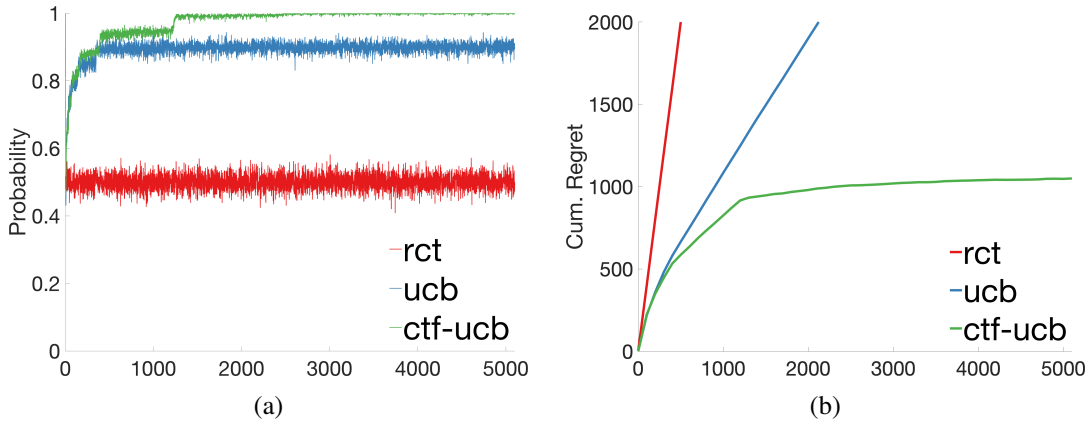
where the bonus function $b_t^{(h)}$ is defined as

$$b_t^{(h)}(s, x, x') = 7H\sqrt{\frac{\ln\left(5|\mathscr{D}(S) \times \mathscr{D}(X) \times \mathscr{D}(X)|T/\delta\right)}{N_t(s, x, x')}} \tag{425}$$

7: Let $V_t^{(h)}(s, x) = \max_{x'} Q_t^{(h)}(s, x, x')$.

8: **end for**

Similarly to the MAB setting, the above proposition implies that there is an MDP environment such that for any online algorithm following Fisherian randomization, it suffers at least a constant regret on average per every step of the interaction. Therefore, these existing algorithms are generally incapable of obtaining an optimal counterfactual policy in MDPs while achieving a sublinear regret. Fortunately, one could augment these RL algorithms with counterfactual reasoning by replacing standard interventions with counterfactual interventions. The proposed `Ctf-UCBVI` (Alg. 10) demonstrates this augmentation procedure in the `UCBVI` algorithm. The following simulation demonstrates the performance of `Ctf-UCBVI` in a simple MDP environment.



(a) (b)

Figure 47: Performance of standard `UCBVI` performing atomic interventions and the augmented `Ctf-UCBVI` using counterfactual interventions.

Environment	Structural Assumptions	Optimality		Autonomy
		Π_{EXP}	Π_{CTF}	
MAB	NUC	✓	✓	✓
	-	✗	✓	✗
MDP	NUC	✓	✓	✓
	-	✗	✓	✗

Table 22: The performance of counterfactual policies Π_{EXP} and experimental policies Π_{CTF} in canonical environments including MABs and MDPs.

Experiment 10 *We evaluate standard `UCBVI` that attempts to maximize rewards based on interventional transitional probabilities $\mathcal{T}_{exp}$ and reward function $\mathcal{R}_{exp}$, while it ignores the agent's intended actions. We also evaluate the augmented `Ctf-UCBVI` algorithm described in Alg. 10 which attempts to maximize cumulative reward based on counterfactual transitional probabilities $\mathcal{T}_{ctf}$ and reward function $\mathcal{R}_{ctf}$. It actively accounts for the agent's intended actions by performing counterfactual interventions. All reported simulations are partitioned into rounds of $T = 5000$ trials averaged over $N = 1000$ repetitions.*

The detailed parameterization of the MDP environment with unobserved confounders is provided in Eq. 5 where the decision horizon $H = 10$. Fig. 47 shows the cumulative regret and probability of selecting optimal arms for evaluated algorithms. Simulation results support our proposed online RL algorithms using counterfactual interventions. Analyses revealed that standard `UCBVI` suffered from a linear regret. Meanwhile, the augmented `Ctf-UCBVI` achieved a sublinear regret, showing that it is able to obtain an optimal counterfactual policy actively accounting for the agent's intended actions in the decision-making process. ■

7.3 The Tradeoff between Autonomy and Optimality

We have established that an agent following a counterfactual policy generally outperforms an interventional one, as summarized in Table 22. The "Environment" column lists canonical decision-making environments covered in this section, such as MABs and MDPs. The "Structural Assumptions" column details additional structural assumptions applied to the environment, including no unmeasured confounding (NUC, Def. 13). The "Optimality" column indicates whether optimizing within the corresponding policy space could lead to an optimal decision strategy, using an optimal counterfactual policy in Π_{CTF} as the baseline. Here, a check (cross) mark in Π_{EXP} under "Optimality" denotes whether experimental policies are capable (or not) of achieving optimal performance in the respective environment, compared to their counterfactual counterparts. In environments where unobserved confounders generally exist, full autonomy (as indicated in the "Autonomy" column) is attainable only when it does not compromise optimality. This occurs when the performance of experimental policies Π_{EXP} and counterfactual policies Π_{CTF} coincide.

This suggests a fundamental tradeoff between optimality and autonomy in the design of RL systems. While full autonomy is preferable, the agent could potentially achieve superior performance by leveraging a human's capabilities through counterfactual reasoning. For instance, as demonstrated in Example 54, a counterfactual policy in Π_{CTF} characterizes interactions in a semi-autonomous system that incorporates the human operator's intended action as input. On the other hand, deploying an interventional policy in Π_{EXP} eliminates the need for a human operator in the underlying environment, resulting in a fully autonomous system. Consequently, Thm. 16 implies that while full autonomy is preferable, the agent could potentially achieve better performance by leveraging the human's intuition through a counterfactual decision criterion.

We will model this autonomy-optimality tradeoff as a constrained transfer of control (TOC) problem such that a decision system tries to maximize its rewards while repeatedly switching between experimental and counterfactual policies, subject to a budget constraint over the total time of using the agent's intended action, i.e., no more than $\delta \in (0, 1)$ ratio of the total running time. For instance, an autonomous vehicle could ask for the human driver's input when necessary, but no more than $\delta = 10\%$ of the total expected driving time. Formally, we first define hybrid policies, which is a restricted family of counterfactual policies Π_{CTF} that contains experimental policies Π_{EXP}.

Definition 27 (Hybrid Policies) *For an MDP environment $\mathcal{M}^*$, a hybrid policy space Π_{HYB} is a subset of counterfactual policies Π_{CTF}. For any hybrid policy $\pi = (\pi_1, \ldots, \pi_H)$ in Π_{HYB}, its decision rule $\pi_i = (f_i \circ g_i)$ is a composition of functions f_i, g_i such that, for $i = 1, \ldots, H$,*

- *g_i is a probability distribution mapping from the current state S_i to an extended action $A_i \in \{0, 1\}$ where "0" stands for* experimental decision criterion *and "1" for* counterfactual decision criterion*. That is,*

$$A_i \sim g_i\left(A_i \mid S_i\right) \tag{429}$$

- *f_i is a probability distribution mapping from the current state S_i, the intended action X_i, and the extended acton A_i to the realized action X_i'. Moreover, the extended action A_i decides the realized action X_i as follows:*

$$f_i\left(X_i \mid S_i, X_i', A_i\right) = \begin{cases} f_i\left(X_i \mid S_i\right), & \text{if } A_i = 0 \\ f_i\left(X_i \mid S_i, X_i'\right), & \text{if } A_i = 1 \end{cases} \tag{430}$$

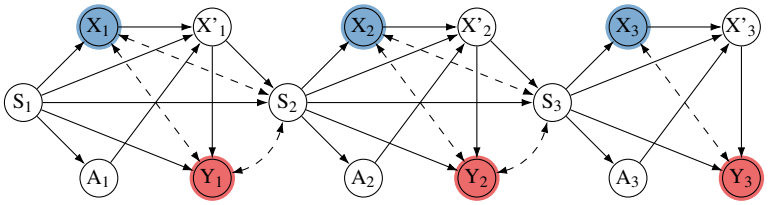


Figure 48: Causal diagram for the MDP environment induced by a hybrid policy where extended actions A_i are added to control the modes of experimental and counterfactual decision criteria.

■

In words, the decision-making process of a hybrid policy consists of two phases. First, it determines an extended action $A_i \in \{0, 1\}$ based on the current state S_i. Here, A_i is a switch variable indicating whether to utilize counterfactual reasoning ($A_i = 1$), or stay in the standard experimental decision criterion ($A_i = 0$). The human's intuition X_i is included as an evidence to decide the realized decision X_i' if $A_i = 1$; otherwise, it is ignored. Fig. 48 shows the augmented causal diagram of an MDP environment induced by a hybrid policy where extended actions A_i are added to control the modes of decision-making criterion. Fix a budget $\delta \in [0, 1]$. For a hybrid policy π that does not apply counterfactual decision criterion with a probability higher than δ as the decision stage i grows, it must satisfy the following constraint:

$$\lim_{i \to \infty} P_\pi \left(A_i = 1 \right) \leq \delta. \tag{431}$$

If the above equation holds, this means that an agent following policy π will not utilize the human's intuition more than $\delta \times 100\%$ of the total running time in long-term.

Example 57 *Recall the MDP environment $\mathcal{M}^*$ described in Eq. 5, where the decision horizon $H = \infty$. Let a hybrid policy $\pi = (\pi_1(X_1 \mid X_1, S_1), \pi_2(X_2 \mid X_2, S_2), \dots)$ such that for every $i = 1, 2, \dots$, the decision rule π_i is defined as,*

$$\pi_i \triangleq X_i \leftarrow \neg X_i' \cdot A_i + \neg S_i \cdot (1 - A_i) \tag{432}$$

where the extended action A_i is drawn over the binary domain $\{0, 1\}$ such that $P(A_i = 1) = 0.1$. Evidently, an agent following such a hybrid policy π must satisfy $P_\pi \left(A_i = 1 \right) = 0.1$ for every $i = 1, 2, \dots$. That is, the agent will not utilize the intended action for more than $0.1 \times 100\%$ of total running time in long term. ■

We next introduce a planning algorithm to solve for an optimal hybrid policy in an MDP environment subject to the budget constraint in Eq. 431. Our previous discussion provided dynamic programming approaches (e.g., value iteration and policy iteration) for optimizing the discounted expected cumulative rewards over counterfactual policies. Here, we first describe an alternative planning strategy using linear programming. Formally, optimizing discounted rewards over counterfactual policies Π_{CTF} in an MDP environment $\mathcal{M}^*$ can be reduced to solving the following equiv-

alent linear program (LP) (d'Epenoux, 1963; Kallenberg, 1983),

$$\begin{aligned} \max \quad & \sum_{s,x,x'} \mathcal{R}_{\text{ctf}}(s,x,x')\phi(s,x,x') \\ \text{subject to} \quad & \forall s,x \in \mathcal{S}\times\mathcal{X}, \quad \phi(s,x,x') \geq 0 \\ & \sum_{x'} \phi(s,x,x') = \alpha\left(s,x\right) + \gamma \sum_{s',x'',x'} \phi\left(s',x'',x'\right) \mathcal{T}_{\text{ctf}}(s,x,x',s',x'') \end{aligned} \tag{433}$$

where $\alpha\left(s,x'\right) = P(S^{(1)}=s, X^{(1)}=x')$ specifies the observational distribution over the initial state and action. The optimization variables $\phi(s,x,x')$ are called the *occupation measure* of a policy, where $\phi(s,x,x')$ is the total discounted number of times action $X_i = x$ is realized in the observed state $S_i = s$, provided with the intended action $X_t = x$. An optimal counterfactual policy in Π_{CTF} is stationary and can be computed from a solution to the above LP as, for $i = 1, 2, \ldots,$

$$\pi_i^*(x'|s,x) = \frac{\phi(s,x,x')}{\sum_{x'}\phi(s,x,x')}. \tag{434}$$

Next we extend the LP formulation in Eq. 433 to solve for an optimal hybrid policy under a budget constraint. Let optimization variables $\phi(s,x,x',a)$ denote the *occupation measure* of a hybrid policy over the realized action $X_i' = x'$, observed state $S_i = s$, the intended action $X_t = x$, and the extended action $A_i = a$. Since the extended action A_i does not directly affect the reward signal Y_i and next state S_{i+1}, the transition probabilities $\mathcal{T}_{\text{ctf}}$ and reward function $\mathcal{R}_{\text{ctf}}$ induced by hybrid policies remain the same as those induced by counterfactual policies. An unconstrained hybrid policy optimizing the MDP environment is thus obtainable by solving the following LP,

$$\begin{aligned} \max \quad & \sum_{s,x,x',a} \mathcal{R}_{\text{ctf}}(s,x,x')\phi(s,x,x',a) \\ \text{subject to} \quad & \forall s,x \in \mathcal{S}\times\mathcal{X}, \forall a \in \{0,1\}, \quad \phi(s,x,x',a) \geq 0 \\ & \sum_{x',a} \phi(s,x,x',a) = \alpha\left(s,x\right) + \gamma \sum_{s',x'',x',a} \phi\left(s',x'',x',a\right) \mathcal{T}_{\text{ctf}}(s,x,x',s',x'') \end{aligned} \tag{435}$$

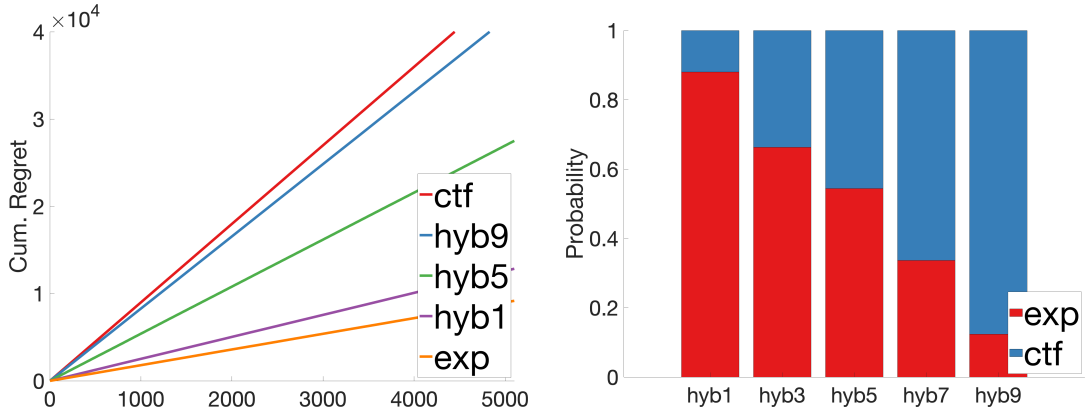
Meanwhile, the budget constraint over the intended action in Eq. 431 could be written as:

$$\sum_{s,x,x'} \phi\left(s,x,x',1\right) \leq \delta \sum_{s,x,x',a} \phi\left(s,x,x',a\right) \tag{436}$$

$$\frac{\phi\left(s,x,x',0\right)}{\sum_{x'}\phi\left(s,x,x',0\right)} = \frac{\sum_{x}\phi\left(s,x,x',0\right)}{\sum_{x,x'}\phi\left(s,x,x',0\right)} \tag{437}$$

$$\frac{\sum_{x'}\phi\left(s,x,x',a\right)}{\sum_{x',a}\phi\left(s,x,x',a\right)} = \frac{\sum_{x,x'}\phi\left(s,x,x',a\right)}{\sum_{x,x',a}\phi\left(s,x,x',a\right)} \tag{438}$$

Among the above equations, Eq. 436 ensures that for an agent operating in the MDP environment, its total time steps applying counterfactual decision criterion ($A_i = 1$) is no more than $\delta \times 100\%$ of the total time steps (discounted so that future visits count less than present ones). Eq. 437 ensures that when an agent applies the experimental decision criterion ($A_i = 0$), the policy π does not take the intended action $X_i' = x'$ as an input. Finally, Eq. 438 reflects the functional constraint



(a) (b)

Figure 49: Simulations comparing the performance (a) and occupancy composition (b) of *exp* and *ctf* decision criteria; y-axis in (b) represents the ratio of total time performing the experimental or counterfactual decision criterion.

that the extended action A_i only depends on the current state S_i. An optimal hybrid policy in Π_{HYB} satisfying the δ-budget constraint is thus obtainable by solving the LP specified in Eq. 435 subject to additional constraints in Eqs. 436 - 438. This mathematical program forms a polynomial optimization problem (Tuy et al., 1998), which is neither linear nor convex. Despite its difficulty, several efficient methods of polynomial optimization can be used in this case, for example, the RLT method (Sherali and Adams, 2013), and a SDP relaxation method (Lasserre, 2001).

Experiment 11 *We evaluate the performance of optimal hybrid policies in an MDP environment subject to different budget constraints. Recall that δ is a constraint over the ratio between the total time of performing the counterfactual decision criterion (using the intended action) and the total running time of the system. We compute policies for three hybrid agents with the ratio constraint δ set to $0.1, 0.5, 0.9$, labeled as* hyb1, hyb5 *and* hyb9*, respectively. We also include the best experimental and counterfactual policies in Π_{EXP} and Π_{CTF} as the baseline, labeled as* exp *and* ctf *respectively. We use value iteration for MDP planning. As for hybrid policy planning, we employ the SDP relaxation method for polynomial optimization. SDPs are constructed using SparsePOP (Waki et al., 2008) and solved with SeDuMi (Sturm, 1999).*

Detailed parameterization of the MDP environment with unobserved confounders is provided in Eq. 5. Fig. 49a shows the discounted cumulative reward for all algorithms. Simulation results reveal that the performance of hybrid policies converges to the best possible counterfactual policy as $\delta \to 1$. In particular, hyb1 *($\delta = 0.1$) shows limited performance improvement over the experimental policy* exp*, while* hyb9 *($\delta = 0.9$) experiences higher cumulative reward, which is comparable to the best counterfactual policy* ctf*. Predictably, the performance of* hyb5 *($\delta = 0.5$) lies in between* hyb1 *and* hyb9*. We also show the composition graph of experimental and counterfactual decision criteria in 49b. Two hybrid policies with $\delta = 0.3, 0.7$ are included. The simulations support that the polynomial optimization reduction worked as expected, where* hyb1 *and* hyb3 *tend to stay in the autonomous mode. In contrast,* hyb7 *and* hyb9 *are more semi-autonomous and actively account for the human's intended action. Unsurprisingly,* hyb5 *kept neutral.* ■

This section investigated a novel interaction regime, called counterfactual randomization, that allows the agent to actively account for human intuition during decision-making using counterfactual reasoning. Our analysis revealed that in almost all cases, a standard agent employing Fisherian randomization is constrained to sub-optimal behaviors; while a counterfactual agent is able to consistently achieve better performance. More generally, our results implied that human intuition should be kept "in the loop" as long as it has access to information about the tasks at hand, even after the agent completes its learning and builds a model of the environment. To resolve the tension between the autonomy and optimality of the system, we proposed a novel RL task subject to a budget constraint. Automated decision-making systems are playing an increasingly prominent role in society, and we hope this work constitutes a step towards a better understanding of the principles underlying human-machine interactions.

8. Causal Imitation Learning (CRL Task 4)

Reinforcement Learning (RL) has been deployed and shown to perform exceptionally well in highly complex environments in the past decades (Sutton and Barto, 1998; Mnih et al., 2013; Silver et al., 2016; Berner et al., 2019; Kumar et al., 2022). One critical assumption behind many of the classical RL algorithms is that the reward function could be well-specified. In many real-world applications, however, it might be impractical to design a suitable reward function that evaluates each and every scenario (Randløv and Alstrøm, 1998; Ng et al., 1999). For example, in the context of human driving, it is challenging to design a precise reward function, and experimenting in the environment could be ill-advised; still, watching expert drivers operate is usually feasible.

In the context of reinforcement learning, the *imitation learning* (IL) paradigm investigates the problem of how an agent should behave and learn in an environment with an unknown reward function by observing demonstrations from a human expert (Argall et al., 2009; Billard et al., 2008; Hussein et al., 2017; Osa et al., 2018). Formally, a causal imitation learning task is characterized by the following signature.

$$\mathcal{T}_{\text{imitate}} = \left\langle \mathcal{I} = \text{see}, \mathcal{A} = \mathcal{G}, \Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}, \mathcal{R} = \emptyset \right\rangle. \tag{439}$$

This means that the agent will try to find a policy π^* such that

$$\pi^* = \arg\max_{\pi \in \Pi_{\text{EXP}}} \mathbb{E}_{\pi}^{\mathcal{M}^*} \left[\mathcal{R}(\boldsymbol{Y}) \middle| \mathcal{G}, \mathcal{D}_{\text{obs}} \sim P(\boldsymbol{V}), \mathcal{R} = \emptyset \right], \tag{440}$$

the distinct feature of the task is that the reward function measuring the system's performance is not fully specified and is unknown from the learner's perspective.

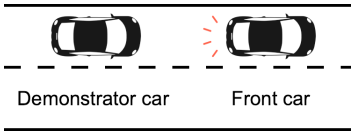


Figure 50: The tail light of the front car is unobserved in highway (aerial) drone data.

An imitation learning agent attempts to learn a policy in space Π from the demonstration data generated by a demonstrator (e.g., a driver, a physician), following a different behavioral policy. For every episode $t = 1, \ldots, T$, the agent observes the demonstrator operating in the underlying environment $\mathcal{M}^*$, and receives trajectories $\boldsymbol{V}^{(t)} \sim P(\boldsymbol{V})$ drawn from the observational distribution. It could also access certain structural assumptions of the environment, e.g., a causal diagram $\mathcal{A} = \mathcal{G}$. Compared with off-policy learning and causal identification tasks, the departing point of imitation learning is that the reward function is not revealed to the learning agent and is not well-specified ($\mathcal{R} = \emptyset$), posing a significant learning challenge.

Corollary 6 *Let endogenous variables $\boldsymbol{X}, \boldsymbol{Y} \subseteq \boldsymbol{V}$. If detailed parametrization of the reward function $\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}$ is unknown, the expected reward $\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})]$ for any policy π over actions $\boldsymbol{X}$ is not identifiable from a causal diagram $\mathcal{G}$.*

The following example demonstrates non-identifiability posed by an unknown reward function.

Example 58 *For concreteness, consider a learning scenario depicted in Fig. 50, describing trajectories of human-driven cars collected by drones flying over highways (Krajewski et al., 2018; Etesami and Geiger, 2020). Using such data, we want to learn a driving policy $\pi(x)$ deciding on the acceleration (action) X of the following car to optimize the distance Y between the following car and the front car. In reality, the human demonstrator's behaviors are affected by an unobserved noise U representing the operational error. The driver's performance is evaluated by an unknown polynomial reward function $\mathcal{R}(Y)$ over the car distance Y.*

More specifically, consider an MAB environment $\mathcal{M}^*$ described by the tuple

$$\mathcal{M}^* = \langle \boldsymbol{U} = \{U\}, \boldsymbol{V} = \{X, Y\}, \mathscr{F}, P(\boldsymbol{U}) \rangle \tag{441}$$

where the structural functions $\mathscr{F}$ are given by:

$$\mathscr{F} = \begin{cases} X \leftarrow U, \\ Y \leftarrow X \end{cases} \tag{442}$$

$U \in \{0, 1\}$ is a binary variable drawn from the exogenous distribution $P(U = 1) = 0.9$. Due to the uncertainty of the reward function $\mathcal{R}(Y)$, the expected reward $\mathbb{E}_\pi[\mathcal{R}(Y)]$ for any driving policy $\pi(X)$ is not identifiable from the observational distribution $P(X, Y)$.

To make this argument more precise, let $\mathcal{R}(Y) = \alpha Y$ be a linear function with an unknown real coefficient $\alpha \in \mathbb{R}$. For any policy $\pi(X)$, the expected reward $\mathbb{E}_\pi[\mathcal{R}(Y)]$ is given by

$$\begin{aligned} \mathbb{E}_\pi[\mathcal{R}(Y)] &= \alpha \mathbb{E}_{X \leftarrow 0}[Y]\, \pi(X = 0) + \alpha \mathbb{E}_{X \leftarrow 1}[Y]\, \pi(X = 1) && (443) \\ &= \alpha \pi(X = 1) && (444) \end{aligned}$$

The last step holds since values of Y are determined by $Y \leftarrow X$. Note that every coefficient $\alpha \in \mathbb{R}$ defines a unique expected reward $\mathbb{E}_\pi[\mathcal{R}(Y)]$. Since α is not a parameter of the SCM $\mathcal{M}^*$, changing values of α does not affect the evaluation of the observational distribution $P(X, Y)$. This means the expected reward $\mathbb{E}_\pi[\mathcal{R}(Y)]$ is not uniquely discernible from the observational distribution $P(X, Y)$ in MAB models, i.e., $\mathbb{E}_\pi[\mathcal{R}(Y)]$ is not identifiable if the reward function $\mathcal{R}$ is unknown. ■

Corol. 6 implies that when the reward function $\mathcal{R}$ is unknown, it is infeasible to uniquely determine the expected reward $\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})]$ from the observational distribution $P(\boldsymbol{V})$ in the causal diagram $\mathcal{G}$. This precludes direct applications of causal identification approaches described in Sec. 4.3, including do-calculus learning (Pearl, 2000), `Identify` algorithm (Tian, 2002), and soft-do-calculus learning (Correa and Bareinboim, 2020a). To circumvent issues of non-identifiability, a common approach is to assume that the observed trajectories are generated by an "expert" demonstrator with satisfactory performance $\mathbb{E}[\mathcal{R}(\boldsymbol{Y})]$, e.g., no less than a certain threshold ($\mathbb{E}[\mathcal{R}(\boldsymbol{Y})] \geq \tau$). If we could find a policy π that performs at least as well as the expert's policy, the agent's performance is also guaranteed to be satisfactory.

Definition 28 *For a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$, an* imitating policy π^* *is a policy such that its expected reward is lower bounded by the expert's reward, i.e.,*

$$\underbrace{\mathbb{E}_{\pi^*}\left[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*\right]}_{\textit{Agent's Performance}} \geq \underbrace{\mathbb{E}\left[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*\right]}_{\textit{Expert's performance}} \tag{445}$$

■

In words, the right-hand side represents the expert's performance that the agent wants to achieve, while the left-hand side represents the real expected reward experienced by the agent evaluated in the underlying environment $\mathcal{M}^*$. We will call this the *fundamental equation of the imitation learning problem*, which the agent aims to solve toward finding a policy π^* that could perform at least as well as the demonstrating expert.

The literature can be partitioned into two major learning modalities that realize imitation :

- *behavioral cloning* (BC) (Widrow, 1964; Pomerleau, 1989; Muller et al., 2006; Mülling et al., 2013; Mahler and Goldberg, 2017), and
- *inverse reinforcement learning* (IRL) (Ng et al., 2000; Ziebart et al., 2008; Ho and Ermon, 2016; Fu et al., 2017).

Specifically, BC methods attempt to directly mimic the expert's behavior policy by learning a mapping from the observed states to the expert's action via supervised learning. On the other hand, IRL methods first learn a surrogate reward function under which the expert's behavior policy is optimal. The imitator then obtains a policy using standard off-policy learning methods (see Sec. 4.1) to maximize the learned reward function. Under some common assumptions, both BC and IRL can obtain policies that achieve the expert's performance (Ng et al., 2000; Abbeel and Ng, 2004). When additional parametric knowledge about the reward function is provided, IRL may produce a policy that outperforms the expert's in the underlying environment (Syed and Schapire, 2008; Li et al., 2017; Yu et al., 2020).

Despite the performance guarantees provided by existing BC and IRL methods, these are contingent on the assumption that the expert's input observations match those available to the imitator. On the other hand, when some expert's observed states remain latent to the imitator, unobserved confounders (UCs) are generally present in the demonstration data, violating the NUC assumption (Def. 13). Perhaps surprisingly, we will show later in this section, when the NUC does not hold, naively applying BC or IRL methods does not necessarily lead to satisfactory performance, even though the expert itself behaves optimally. After all, it is unclear how to perform imitation learning with unobserved confounding in the expert's demonstrations. This section answers this question and, more broadly, investigates the problem of imitation learning through causal lenses. We will provide novel algorithms capable of learning an imitating policy that performs at least as well as the expert from the demonstration data while allowing the presence of UCs. In particular, our contributions are summarized as follows.

- **Confounding Robust BC.** Sec. 8.1 introduces a sufficient and necessary graphical criterion for determining the feasibility of BC-type learning procedure from demonstration data and qualitative knowledge about the data-generating process represented as a causal diagram. When such a condition holds, an imitating policy is obtainable using standard BC algorithms to achieve the expert's performance.

- **Confounding Robust IRL.** Sec. 8.2 derives a new graphical condition for deciding whether an imitating policy can be computed from the available data and knowledge, which provides a robust generalization of current IRL algorithms to general settings where the NUC assumption does not hold. These algorithms include GAIL (Ho and Ermon, 2016), and MWAL (Syed and Schapire, 2008).

8.1 Causal Behavioral Cloning

We first consider the behavioral cloning approach, where the agent attempts to learn an imitating policy by mimicking conditional observational distributions $P(X_i \mid \boldsymbol{Z}_i)$ over domains of every action $\boldsymbol{X}_i$ given a subset of input states $\boldsymbol{Z}_i \subseteq \boldsymbol{S}_i$. Formally

Definition 29 (Behavioral Cloning Policy) *Let $\Pi = \{\langle X_i, \boldsymbol{S}_i\rangle\}_{i=1}^{H}$ be a policy space, and $P(\boldsymbol{V})$ be an observational distribution. A* behavioral cloning policy $\pi \in \Pi$ *is an expression in terms of $P(\boldsymbol{V})$ such that for every $i = 1, \ldots, H$, $\pi_i(X_i \mid \boldsymbol{Z}_i) = P(X_i \mid \boldsymbol{Z}_i)$ for some $\boldsymbol{Z}_i \subseteq \boldsymbol{S}_i$.*[63] ■

There exist algorithms in imitation learning literature to perform behavioral cloning from observed demonstration data (Widrow, 1964; Pomerleau, 1989; Muller et al., 2006; Mahler and Goldberg, 2017). The following example illustrates BC learning in an MAB environment.

Example 59 (Behavioral Cloning in MAB) *Consider again the MAB environment $\mathcal{M}^*$ described in Eq. 442 concerning with learning a driving policy $\pi(X)$ following the front car. Since $\pi(X)$ belongs to a policy space $\Pi = \{\langle X, \emptyset\rangle\}$, a behavioral cloning policy $\pi_{\text{BC}}(X)$ is given by*

$$\pi_{\text{BC}}(X = 1) = P(X = 1) \tag{446}$$
$$= P(U = 1) \tag{447}$$

Computing the above equation gives $\pi_{\text{BC}}(X = 1) = 0.9$. Recall the reward function $\mathcal{R}(Y) \leftarrow \alpha Y$ is a linear function with an unknown coefficient α. Evaluating the expected reward $\mathcal{R}(Y)$ in submodel $\mathcal{M}^*_{\pi_{\text{BC}}}$ gives, following the decomposition in Eq. 444:

$$\mathbb{E}_{\pi_{\text{BC}}}[\mathcal{R}(Y)] = \alpha \pi_{\text{BC}}(X = 1) \tag{448}$$
$$= 0.9\alpha \tag{449}$$

We will next show that the BC policy π_{BC} achieves the demonstrator's performance. By evaluating the expected reward $\mathcal{R}(Y)$ in SCM $\mathcal{M}^*$, we obtain

$$\mathbb{E}[\mathcal{R}(Y)] = \alpha \mathbb{E}[Y] \tag{450}$$
$$= \alpha P(U = 1) \tag{451}$$

Computing the above equation gives the evaluation of the demonstrator's performance $\mathbb{E}[\mathcal{R}(Y)] = 0.9\alpha$, which matches the performance of BC policy π_{BC}. ■

63. There could exist multiple behavioral policies in a policy space Π simulating the same conditional distributions $P(X_i \mid \boldsymbol{Z}_i)$. For instance, for $\Pi = \{\langle X, \emptyset\rangle\}$ and $P(X = 1) = 0.9$, behavioral policies $\pi, \pi' \in \Pi$ are given by $\pi(X = 1) = \mathbb{1}\{U \leq 0.9\}$ and $\pi'(X = 1) = \mathbb{1}\{U \geq 0.1\}$ where U is an uniform distribution over $[0, 1]$.

More generally, behavioral cloning is able to imitate the expert's performance when the NUC assumption holds (Def. 13) and the input states $\boldsymbol{S}_i$ for every action X_i are sufficiently large (to be defined). More precisely, provided with the NUC, for any policy $\pi \in \Pi$, the expected reward $\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})]$ could be decomposed as, following the IPW identification formula (Thm. 2),

$$\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*] = \sum_{\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H} E[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H] P(\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H) \prod_{i=1}^{H} \frac{\pi_i(x_i \mid \boldsymbol{s}_i)}{P(x_i \mid \bar{\boldsymbol{x}}_{i-1}, \bar{\boldsymbol{s}}_i)}. \tag{452}$$

Let π be a behavioral cloning policy in Π such that $\pi_i(X_i \mid \boldsymbol{S}_i) = P(X_i \mid \boldsymbol{S}_i)$ for action X_i. The above equation could be written as

$$\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*] = \sum_{\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H} E[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H] P(\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H) \prod_{i=1}^{H} \frac{P(x_i \mid \boldsymbol{s}_i)}{P(x_i \mid \bar{\boldsymbol{x}}_{i-1}, \bar{\boldsymbol{s}}_i)}. \tag{453}$$

Let the input states $\boldsymbol{S}_i$ for every action X_i be sufficiently large such that the following independence relationships hold in the observational distribution $P(\boldsymbol{V})$,

$$\left(X_i \perp\!\!\!\perp \bar{\boldsymbol{X}}_{i-1}, \bar{\boldsymbol{S}}_{i-1} \mid \boldsymbol{S}_i\right) \ \forall i = 1, \ldots, H. \tag{454}$$

One example for the above condition to hold is when states $\boldsymbol{S}_i$ contain all observed parents $\boldsymbol{PA}_i$ for action X_i. Since the NUC holds, values of every X_i are determined by independent noise $\boldsymbol{U}_i$ given input states $\boldsymbol{S}_i$. Eq. 453 could be further written as

$$\begin{aligned}
\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*] &= \sum_{\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H} E[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H] P(\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H) \prod_{i=1}^{H} \frac{P(x_i \mid \boldsymbol{s}_i)}{P(x_i \mid \boldsymbol{s}_i)} && (455)\\
&= \sum_{\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H} E[\mathcal{R}(\boldsymbol{Y}) \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H] P(\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{s}}_H) && (456)\\
&= \mathbb{E}[\mathcal{R}(\boldsymbol{Y})] && (457)
\end{aligned}$$

The last step follows by summing over states and actions $\bar{\boldsymbol{S}}_H, \bar{\boldsymbol{X}}_H$. In words, the behavior policy $\pi \in \Pi$ achieves the expert's performance. Formally,

Theorem 19 (Behavioral Cloning from NUC) *Let $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$ be a CDM where $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^{H}$ and $\mathcal{R} : \Omega(\boldsymbol{Y}) \mapsto \mathbb{R}$. Consider the following conditions:*

1. *The NUC condition (Def. 13) holds for the policy space Π in SCM $\mathcal{M}^*$;*
2. *For every action X_i, $i = 1, \ldots, H$, its endogenous parents $\boldsymbol{PA}_i \subseteq \boldsymbol{S}_i$.*

Then, there is a behavioral cloning policy $\pi \in \Pi$ (Def. 29), where the expected reward $\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y})]$ matches the expert's performance evaluated in $\mathcal{M}^*$,

$$\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*] = \mathbb{E}[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*]. \tag{458}$$

Moreover, such a policy π is given by $\pi_i(X_1 \mid \boldsymbol{S}_i) = P(X_i \mid \boldsymbol{S}_i)$ for every $i = 1, \ldots, H$. ■

Example 59 shows that behavioral cloning is able to achieve the expert's performance in an MAB environment. Our next example demonstrates behavioral cloning in the sequential setting.

Example 60 (Behavioral Cloning in DTR) *Consider the 2-stage DTR $\langle \mathcal{M}^*, \Pi, Y \rangle$ where SCM $\mathcal{M}^*$ is described by Eq. 78 with coefficients $\alpha_1 = \alpha_2 = 0$; and the policy space $\Pi = \{\langle X_1, \{S_1\}\rangle, \langle X_2, \{S_1, X_1, S_2\}\rangle\}$. Evidently, Condition 1 of Thm. 19 is satisfied since NUC holds in this model (please revisit Example. 25). Also, Condition 2 of Thm. 19 holds since the endogenous parents of the actions X_1, X_2 are $\boldsymbol{PA}_1 = \{S_1\}$ and $\boldsymbol{PA}_2 = \{S_2\}$, respectively, which are both contained in the input states. Applying Thm. 19 implies the learner could achieve the expert's performance with a behavioral cloning policy $\pi = (\pi_1, \pi_2)$ given by*

$$\pi_1(X_1 \mid S_1) = P(X_1 \mid S_1), \qquad \pi_2(X_2 \mid S_1, X_1, S_2) = P(X_2 \mid S_1, X_1, S_2) \tag{459}$$

Evaluating the above equation gives decision rules

$$\pi_i : X_i \leftarrow \mathbb{1}\{3S_i + U_i > 0\}, \ \ \forall i = 1, 2 \tag{460}$$

where U_i, $i = 1, 2$, are independent variables drawn from a logistic distribution $\texttt{Logistic}(0, 1)$. It follows from Eq. 78 submodel $\mathcal{M}^*_\pi$ coincides with SCM $\mathcal{M}^*$. As a consequence, we must have $\mathbb{E}_\pi[Y] = \mathbb{E}[Y]$, i.e., the learned policy achieves the expert's performance. ■

However, the behavioral cloning strategy does not always achieve the expert's performance in all environments, especially when the input variables of the imitator's and the expert's policy mismatch, and unobserved confounders generally exist in the demonstration data. Our next example illustrates the challenges of unobserved confounding.

Example 61 (Behavioral Cloning fails without NUC) *Consider again the driving scenario described in Example 58. Suppose now that the distance Y between the demonstrator and font car is affected by the deceleration U of the front car. Also, the human driver perceives the deceleration U of the front car through its tail light and determines the action X. However, since the tail light is not recorded in the drone footage, variable U is thus unobserved from the imitator's perspective.*

More specifically, this environment is described by an SCM $\mathcal{M}$ defined as

$$\mathcal{M} = \langle \boldsymbol{U} = \{U\}, \boldsymbol{V} = \{X, Y\}, \mathscr{F}, P(\boldsymbol{U}) \rangle \tag{461}$$

where structural functions $\mathscr{F}$ are given by

$$\mathscr{F} = \begin{cases} X \leftarrow \neg U, \\ Y \leftarrow X \oplus U \end{cases} \tag{462}$$

$U \in \{0, 1\}$ is a binary variable drawn from the exogenous distribution $P(U = 1) = 0.5$. Since U is now an observed confounder affecting both action X and outcome Y, the NUC condition does not hold in this environment $\mathcal{M}$.

Let the driver's performance be measured by a reward function $\mathcal{R}(Y) = \alpha Y$ with an unknown coefficient $\alpha \in \mathbb{R}$. Evaluating the expected reward in $\mathcal{M}$ gives:

$$\mathbb{E}[\mathcal{R}(Y)] = \alpha \mathbb{E}[Y] \tag{463}$$
$$= \alpha \mathbb{E}[X \oplus U] \tag{464}$$
$$= \alpha \mathbb{E}[\neg U \oplus U] \tag{465}$$

The last step holds since $X \leftarrow \neg U$ in SCM $\mathcal{M}^*$. Computing the above equation gives the expert's performance $\mathbb{E}[\mathcal{R}(Y)] = \alpha$.

We now apply the behavioral cloning strategy in Thm. 19 and see if it imitates the expert's performance even when the NUC does not hold. Mimicking the marginal distribution $P(X)$ results in a behavioral policy $\pi_{\text{BC}}(X)$ such that

$$\pi_{\text{BC}}(X = 1) = P(X = 1) \tag{466}$$
$$= P(U = 1) \tag{467}$$

Computing the above equation gives $\pi_{\text{BC}}(X = 1) = 0.5$, i.e., the imitator randomly accelerates the demonstrator car. Evaluating the expected reward $\mathcal{R}(Y)$ in submodel $\mathcal{M}_{\pi_{\text{BC}}}$ implies

$$\mathbb{E}_{\pi_{\text{BC}}}[\mathcal{R}(Y)] = \sum_x \alpha \mathbb{E}_x[Y]\, \pi_{\text{BC}}(x) \tag{468}$$
$$= 0.5\alpha \left(\mathbb{E}_{X \leftarrow 0}[Y] + \mathbb{E}_{X \leftarrow 1}[Y]\right) \tag{469}$$

Since values of Y are given by $Y \leftarrow X \oplus U$, we further have

$$\mathbb{E}_{\pi_{\text{BC}}}[\mathcal{R}(Y)] = 0.5\alpha \mathbb{E}[0 \oplus U + 1 \oplus U] \tag{470}$$
$$= 0.5\alpha \tag{471}$$

Suppose the actual coefficient $\alpha > 0$ is positive. The imitator's performance $\mathbb{E}_{\pi_{\text{BC}}}[\mathcal{R}(Y)] = 0.5\alpha$ is far from the expert's performance $\mathbb{E}[\mathcal{R}(Y)] = \alpha$. ■

In words, behavioral cloning does not guarantee to achieve the expert's performance when the NUC condition does not hold, which calls for alternative cloning strategies. We will next study a more generalized imitation setting from the expert's demonstration, provided with a causal diagram encoding the underlying qualitative knowledge about the environment.

8.1.1 BACKDOOR CRITERION FOR IMITATION

Our discussion starts with a variant of the sequential backdoor criterion (Def. 15) that allows the learner to imitate the expert's performance (Zhang et al., 2020; Kumor et al., 2021). Recall that for any policy $\pi \in \Pi$, $\mathcal{G}_{\pi_{i+1},\dots,\pi_H}$, $i = 0, \dots, H-1$, is a manipulated graph obtained from the causal diagram $\mathcal{G}$ by replacing incoming arrows of every action node $X_j \in \{X_{i+1}, \dots, X_H\}$, with arrows from input states in $\boldsymbol{S}_j$ to X_j. In the context of imitation learning, $\mathcal{G}_{\pi_{i+1},\dots,\pi_H}$ can be seen as $\mathcal{G}$ with all future actions after the i-th stage of the intervention is already encoded in the graph. Formally, the imitation backdoor criterion is defined as follows:

Definition 30 (Imitation Backdoor Condition) *Let $\mathcal{G}$ be a causal diagram and $\boldsymbol{X}, \boldsymbol{Y} \in \boldsymbol{V}$ be subsets of variables. A policy space $\Pi = \{\langle X_i, \boldsymbol{S}_i \rangle\}_{i=1}^H$ is said to satisfy the* imitation backdoor condition *w.r.t. $\boldsymbol{Y}$ in $\mathcal{G}$ (for short, Π is imitation admissible) if for every policy $\pi \in \Pi$, every action $X_i \in \boldsymbol{X}$, one of the following conditions hold:*

1. *X_i is not an ancestor of $\boldsymbol{Y}$ in $\mathcal{G}_{\pi_{i+1},\dots,\pi_H}$, i.e., $X \notin An(\boldsymbol{Y})_{\mathcal{G}_{\pi_{i+1},\dots,\pi_H}}$;*
2. *$\boldsymbol{S}_i$ d-separates all backdoor path from node X_i to nodes in $\boldsymbol{Y}$ in $\mathcal{G}_{\pi_{i+1},\dots,\pi_H}$, i.e., $(\boldsymbol{Y} \perp\!\!\!\perp X_i | \boldsymbol{S}_i)$ in $\mathcal{G}_{\underline{X_i},\pi_{i+1},\dots,\pi_H}$.* ■

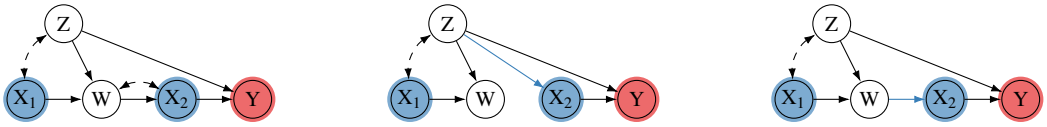


Figure 51: A causal diagram and its manipulated subgraphs.

The first condition in Def. 30 corresponds to the case where an action at X_i does not affect the value of $\boldsymbol{Y}$ once future actions are taken. Since $\mathcal{G}_{\pi_{i+1},\ldots,\pi_H}$ has modified parents for future actions $\bar{\boldsymbol{X}}_{i+1:H}$, the value of X_i might no longer be relevant at all to Y, i.e. Y would get the same input distribution no matter what policy is chosen for X_i. This allows X_i to fail Condition (2), meaning that it is not clonable by itself, but still be part of a clonable set $\boldsymbol{X}$, because future actions can shield Y from errors made at X_i. The second condition is similar to the backdoor criterion where $\boldsymbol{Z}_i$ is a set of variables that effectively encodes all information relevant to imitating X_i with respect to Y. In other words, if the joint distribution $P(\boldsymbol{Z}_i, X_i)$ over the observed states $\boldsymbol{Z}_i$ and action X_i matches when both expert and imitator are acting, then an adversarial reward function Y cannot distinguish between the two and imitation could be successfully realized.

Example 62 (Imitation Backdoor $\not\Rightarrow$ Sequential Backdoor) *We will illustrate the distinction between Conditions (1) and (2) in the causal diagram $\mathcal{G}$ of Fig. 51a. Consider a policy space*

$$\Pi_1 = \{\langle X_1, \emptyset\rangle, \langle X_2, \{Z\}\rangle\}\,. \tag{472}$$

For every policy $(\pi_1, \pi_2) \in \Pi_1$, the manipulated diagram $\mathcal{G}_{\pi_2}$ is shown in Fig. 51b. As for action X_1, since its input states $\boldsymbol{S}_1 = \emptyset$, there is no valid adjustment set that can d-separate X_1 from Y. However, since the policy for action X_2 uses Z as input instead of W or X_1 (i.e. $\pi_2(X_2 \mid Z)$), X_1 will no longer be an ancestor of Y in $\mathcal{G}_{\pi_2}$ and Condition (1) holds. In effect, the action made at X_2 ignores the mistakes made at X_1 due to not having access to unobserved confounders when taking action. Conditioning on covariate node Z d-separates all backdoor paths between X_2 and Y in the subgraph $\mathcal{G}$, satisfying Condition (2). Therefore, the policy space Π_1 is imitation admissible w.r.t. the reward signal Y in the causal diagram $\mathcal{G}$. ■

An interesting observation follows from the above example. While $\Pi_1 = \{\langle X_1, \emptyset\rangle, \langle X_2, \{Z\}\rangle\}$ is imitation admissible in Fig. 51a, Π_1 does not satisfy the sequential backdoor condition of Def. 15. This is the case since the input state $\boldsymbol{S}_1 = \emptyset$ fails to block the backdoor path between X_1 and Y via covariate Z. On the other hand, in some settings, there exist policy spaces satisfying the sequential backdoor condition (Def. 15) but are not imitation admissible.

Example 63 (Sequential Backdoor $\not\Rightarrow$ Imitation Backdoor) *Consider the causal diagram $\mathcal{G}$ described in Fig. 51a and a policy space $\Pi_2 : \{\langle X_1, \{Z\}\rangle, \langle X_2, \{W\}\rangle\}$. For every policy $(\pi_1, \pi_2) \in \Pi_2$, the manipulated diagram $\mathcal{G}_{\pi_2}$ is shown in Fig. 51c. As for action X_1, Condition (2) holds since policy $\pi_1(X_1 \mid Z)$ for X_1 takes Z as input, and conditioning on Z d-separates all backdoor paths from X_1 to reward Y in $\mathcal{G}_{\pi_2}$. As for action X_2, Condition (1) does not hold since X_2 is a direct parent of Y. Condition (2) fails to apply since input variable W is a collider and conditioning on W opens the backdoor path between X and Y in $\mathcal{G}$, e.g., $X_2 \dashleftarrow\dashrightarrow W \leftarrow Z \rightarrow Y$. Still, conditioning on all past actions and states' history X_1, Z, W d-separates backdoor paths from X_2 to Y*

in $\mathcal{G}$. This means Π_2 satisfies the sequential backdoor condition of Def. 15 and the expected reward of policy $\pi \in \Pi_2$ is identifiable from $P(X_1, Z, W, X_2, Y)$. ■

The imitation backdoor condition provides an effective algorithm for deciding whether a policy compatible with a policy space in a causal diagram is imitable or not. Next, we describe some necessary notations.

Definition 31 (Policy Subspace) *For a policy space $\Pi = \{\langle X_1, \boldsymbol{S}_i\rangle\}_{i=1}^{H}$, a policy subspace Π' of Π, denoted by $\Pi' \subseteq \Pi$, is a sequence $\{\langle X_i, \boldsymbol{Z}_i\rangle\}_{i=1}^{H}$ where $\boldsymbol{Z}_i \subseteq \boldsymbol{S}_i$ for every action $X_i \in \boldsymbol{X}$.* ■

In words, every policy space Π is also a subspace of itself. A subspace Π' contained in Π is *proper* if $\Pi' \neq \Pi$ is not equal to Π, which we denote by $\Pi' \subset \Pi$. Note that for every policy $\pi' \in \Pi'$ compatible with a subspace Π', one could always simulate it using a policy $\pi \in \Pi$ such that $\pi_i(x_i \mid \boldsymbol{s}_i) = \pi'_i(x_i \mid \boldsymbol{z}_i)$, $i = 1, \ldots, H$, for all realizations $x_i, \boldsymbol{s}_i, \boldsymbol{z}_i$. That is, input states in the set difference $\boldsymbol{S}_i \setminus \boldsymbol{Z}_i$ do not affect values of action X_i. It follows that policy $\pi' \in \Pi$ is also compatible with space Π if it is compatible with a subspace $\Pi' \subseteq \Pi$.

Theorem 20 (Behavioral Cloning from Imitation Backdoor) *Let $\mathcal{G}$ be a causal diagram, Π be a policy space over actions $\boldsymbol{X}$, and $\boldsymbol{Y} \subseteq \boldsymbol{V}$ be a subset of variables. If there exists a subspace $\Pi' = \{\langle X_i, \boldsymbol{Z}_i\rangle\}_{i=1}^{H}$ contained in Π such that Π' is imitation admissible w.r.t. $\boldsymbol{Y}$ in $\mathcal{G}$, then there is a behavior cloning policy $\pi \in \Pi'$ such that for any SCM $\mathcal{M}^*$ compatible with diagram $\mathcal{G}$,*

$$\mathbb{E}_{\pi}\left[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*\right] = \mathbb{E}\left[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*\right]. \tag{473}$$

Moreover, such a policy $\pi \in \Pi'$ is given by $\pi_i\left(X_i|\boldsymbol{Z}_i\right) = P\left(X_i|\boldsymbol{Z}_i\right)$ for every $i = 1, \ldots, H$. ■

Thm. 20 implies that whenever an imitation admissible subspace $\Pi' \subseteq \Pi$ is found, the expert's performance is achievable using behavioral cloning, i.e., mimicking the conditional distribution $P(X_i|\boldsymbol{Z}_i)$ for every action $X_i \in \boldsymbol{X}$. Moreover, it has been shown that the imitation backdoor criterion is also necessary for determining the feasibility of behavioral cloning for a general class of policy spaces such that for every action X_i, the input states $\boldsymbol{S}_i$ contain all variables preceding X_i following a temporal ordering in the diagram $\mathcal{G}$ (Zhang et al., 2020; Kumor et al., 2021).[64] That is, if there is no imitation admissible subspace in Π, then for any behavioral cloning policy $\pi \in \Pi$, one could always construct an SCM $\mathcal{M}$ compatible with the causal diagram $\mathcal{G}$ such the BC policy π fails to achieve the expert's performance.

Example 64 *Consider again the causal diagram $\mathcal{G}$ described in Fig. 51a and a policy space $\Pi = \{\langle X_1, \{Z\}\rangle, \langle X_2, \{Z, W\}\rangle\}$. Note that $\Pi_1 = \{\langle X_1, \emptyset\rangle, \langle X_2, \{Z\}\rangle\}$ is a subspace contained Π and, as discussed previously, is imitation admissible in diagram $\mathcal{G}$. It follows from Thm. 20 that the expert's performance $\mathbb{E}\left[\mathcal{R}(Y)\right]$ is achievable from observational data using a behavioral cloning policy $\pi = (\pi_1, \pi_2)$ given by $\pi_1(X_1) = P(X_1)$ and $\pi_2(X_2 \mid Z) = P(X_2 \mid Z)$.* ■

For every action $X_i \in \boldsymbol{X}$, the imitation backdoor criterion requires that the covariates $\boldsymbol{Z}_i$ is a back-door adjustment set in the manipulated diagram $\mathcal{G}_{\pi_{i+1}, \ldots, \pi_H}$. There exist efficient methods for finding adjustment sets in the literature (van der Zander and Liśkiewicz, 2020). The learner could run these algorithms on each action X_i iteratively to find each backdoor admissible set $\boldsymbol{Z}_i$ following a reverse topological ordering $X_H \succ X_{H-1} \succ \cdots \succ X_1$, which will lead to an imitation

64. Indeed, this is the largest possible policy space defined over action $\boldsymbol{X}$ in an SCM.

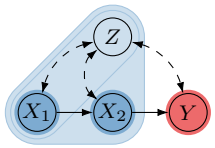
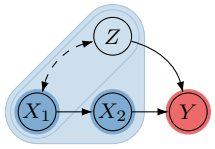
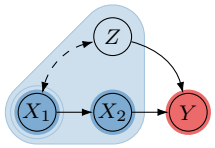
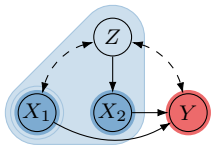
#	Causal Diagram	Causal BC	BC – Observed Parents	BC – All Observed
$\mathcal{G}_1$		$0.04 \pm 0.04\%$	$0.05 \pm 0.04\%$	$\mathbf{0.13 \pm 0.18\%}$
$\mathcal{G}_2$		$0.05 \pm 0.03\%$	$\mathbf{0.20 \pm 0.25\%}$	$0.05 \pm 0.03\%$
$\mathcal{G}_3$		$0.04 \pm 0.03\%$	$\mathbf{0.27 \pm 0.40\%}$	$\mathbf{0.26 \pm 0.39\%}$
$\mathcal{G}_4$		Not Imitable	$\mathbf{0.19 \pm 0.29\%}$	$\mathbf{0.19 \pm 0.29\%}$

Table 23: Performance gap $|\mathbb{E}_{\pi}[Y] - \mathbb{E}[Y]|$ from behavioral cloning using different input states in randomly sampled SCMs consistent with each causal diagram.

admissible subspace $\Pi' = \{\langle X_i, \boldsymbol{Z}_i \rangle\}_{X_i \in \boldsymbol{X}}$ in the end. When the state variables $\boldsymbol{S}_i$, $i = 1, \ldots, H$, contain all variables preceding every action X_i following a topological ordering in the diagram $\mathcal{G}$, (Kumor et al., 2021) provides a polynomial-time algorithm to find an imitation admissible subspace Π'. This means that a policy subspace Π' satisfying the imitation backdoor condition in Def. 30 is generally easier to obtain if it exists. If that is the case, an imitating policy could be obtained from demonstrations by "cloning" the expert's nominal policy, following standard behavioral cloning algorithms (Widrow, 1964; Pomerleau, 1989).

Experiment 12 *We evaluate BC algorithms in randomly sampled SCMs consistent with various causal diagrams. These algorithms use different criteria to select input states/features for every action, which we summarize as follows. (1) Our proposed causal BC selects input states using the imitation backdoor condition, following the procedure described in Thm. 20. (2) Standard BC algorithm mimics the expert's nominal policy, taking observed direct parents for every action as input. (3) Standard BC algorithm considers all state variables available to the imitator at the time of each action, described by the policy space.*

For each causal diagram, $10,000$ *random discrete causal models are sampled, the expert's performance is measured, and then the expert's policy is replaced with imitating policies* $\pi_i(X_i \mid \boldsymbol{Z}_i) = P(X_i|\boldsymbol{Z}_i)$ *for every action* $X_i \in \boldsymbol{X}$*, with input covariates* $\boldsymbol{Z}_i$ *determined by the tested BC algorithm described above. The performance of algorithms is evaluated using the gap between the expert's performance* $\mathbb{E}[Y]$ *and the expected reward of the policy* $\mathbb{E}_{\pi}[Y]$ *obtained by the imitator.*

Simulation results are shown in Table 23, with causal diagrams and policy spaces described in the first column, followed by the performance gap between the expert and the imitator.

For the diagram $\mathcal{G}_1$, including Z when developing a policy for X_1, X_2 leads to a biased answer, which makes the average error of using all observed covariates (red) larger than just the sampling fluctuations present in the other columns. Similarly, Z needs to be considered in $\mathcal{G}_2$, but it is not explicitly used by X_2, so a method relying only on observed parents leads to bias here. In $\mathcal{G}_3$, Z is not observed at the time of determining action X_1, making standard BC algorithms fail to achieve the expert's performance. Our method recognizes that X_2's policy can fix the imitation error made at X_1, and is the only method that leads to an unbiased result. Finally, in $\mathcal{G}_4$, the non-causal approaches cannot determine non-clonability, and return biased results in all such cases. ■

8.2 Causal Inverse RL

This section studies an alternative imitation learning strategy, called *inverse reinforcement learning* (IRL, Ng et al. (2000); Ziebart et al. (2008); Ho and Ermon (2016); Fu et al. (2017)), in a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R} \rangle$. Similar to behavioral cloning, detailed parametrizations of the underlying environment $\mathcal{M}^*$ and the reward function $\mathcal{R}$ are not fully unknown. However, the imitator now has access to a parametric family $\mathscr{R} \subseteq \{\forall \mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}\}$ containing the actual reward function $\mathcal{R}$. We will start the discussion by describing an IRL strategy when the NUC assumption (Def. 13) holds. We will next relax the NUC and study inverse RL in a more general class of causal models, provided with a causal diagram $\mathcal{G}$ encoding its qualitative knowledge.

Following the game-theoretic approach introduced in (Syed and Schapire, 2008), we formulate imitating learning (Def. 28) as learning to play a two-player zero-sum game in which the agent chooses a policy, and the adversarial (e.g., Nature) chooses a worst-case reward function from the parametric reward family $\mathscr{R}$. Now consider the optimization problem defined as follows.

$$\nu^* = \min_{\pi \in \Pi} \max_{\mathcal{R} \in \mathscr{R}} \mathbb{E}\left[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*\right] - \mathbb{E}_\pi\left[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*\right]. \tag{474}$$

The inner maximization in the above equation can be viewed as a *causal IRL* step where we attempt to "guess" a worst-case reward function $\hat{\mathcal{R}} \in \mathscr{R}$ that prioritizes the expert's policy. That is, the gap in the performance between the expert's and the imitator's policies is maximized. Meanwhile, note that the expert's reward $\mathbb{E}\left[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*\right]$ is not affected by the imitating's policy π. The outer minimization is equivalent to a planning step that finds a policy π^* optimizing a CDM $\left\langle \mathcal{M}^*, \Pi, \hat{\mathcal{R}} \right\rangle$ with the worst-case reward $\hat{\mathcal{R}}$. Obviously, the solution π^* is an imitating policy if the performance gap $\nu^* = 0$. In cases where the expert is sub-optimal, we may have $\nu^* < 0$, i.e.,

$$\mathbb{E}\left[\hat{\mathcal{R}}(\boldsymbol{Y}); \mathcal{M}^*\right] < \mathbb{E}_\pi\left[\hat{\mathcal{R}}(\boldsymbol{Y}); \mathcal{M}^*\right], \ \exists \pi \in \Pi \tag{475}$$

In words, the solution π^* will dominate the expert's policy $f_{\boldsymbol{X}}$ in the worst-case scenario, regardless of the detailed form of the reward function $\mathcal{R}$. To some extent, the imitating policy π^* ignores the sub-optimal expert and instead exploits prior knowledge about the unknown reward function. When the prior knowledge is informative, solving the optimization program in Eq. 474 could produce a policy that could significantly outperform the expert in the underlying environment with respect to the unknown reward function, while at the same time guaranteed to be no worse.

Example 65 (Inverse RL in MAB) *Consider again the MAB environment $\mathcal{M}^*$ described in Eq. 442 concerning learning a driving policy from highway footage. Suppose that the reward function $\mathcal{R}(Y) = \alpha Y$ is linear with a positive coefficient $\alpha > 0$. The minimax program in Eq. 474 could be written as*

$$\nu^* = \min_{\pi(X)} \max_{\alpha>0} \alpha\mathbb{E}[Y] - \alpha\mathbb{E}_\pi[Y] \tag{476}$$

Evaluating the distance from the front car Y in the environment $\mathcal{M}^*$ gives:

$$\mathbb{E}[Y] = P(X = 1) = 0.9 \tag{477}$$

Note that the NUC condition holds in this environment. The interventional quantity $\mathbb{E}_\pi[Y]$ is a function of the observational distribution $P(X, Y)$ and policy $\pi(x)$ is given by, following the DP formula in Thm. 3 (or the IPW formula in Thm. 2),

$$\mathbb{E}_\pi[Y] = \mathbb{E}[Y \mid X = 0]\pi(X = 0) + \mathbb{E}[Y \mid X = 1]\pi(X = 1) \tag{478}$$
$$= \pi(X = 1) \tag{479}$$

The last step holds since values of Y are determined by $Y \leftarrow X$. By substituting Eqs. 477 and 479 into Eq. 476, we can further write the performance gap ν^* as:

$$\nu^* = \min_{\pi(X)} \max_{\alpha>0} \alpha\,(0.9 - \pi(X = 1)) \tag{480}$$

For any coefficient $\alpha > 0$, the above program is minimized with a solution $\pi(X = 1) = 1$. Solving the above equation thus leads to an IRL policy $\pi_{\text{IRL}} : X \leftarrow 1$. In this case, the performance gap is equal to $\nu^* = -0.1\alpha < 0$, which means that the IRL imitator outperforms the expert.

To verify this intuition, we evaluate the expected reward $\mathcal{R}(Y)$ in submodel $\mathcal{M}^*_{\pi_{\text{IRL}}}$, following the evaluation formula in Eq. 444,

$$\mathbb{E}_{\pi_{\text{IRL}}}[\mathcal{R}(Y)] = \alpha\pi_{\text{IRL}}(X = 1) \tag{481}$$

This means that the IRL policy achieves the expected reward $\mathbb{E}_{\pi_{\text{IRL}}}[\mathcal{R}(Y)] = \alpha$, which outperforms both the expert and BC's policies $\mathbb{E}[\mathcal{R}(Y)] = \mathbb{E}_{\pi_{\text{BC}}}[\mathcal{R}(Y)] = 0.9\alpha$ (see Example 59 for detailed computations). ■

Despite its clear semantics, solving the optimization problem in Eq. 474 requires the detailed parametrization of the underlying SCM $\mathcal{M}^*$, which is not accessible to the agent in most real-world settings. It is then important to study conditions under which the solution of Eq. 474 is identifiable and could be formulated from the observational distribution $P(\boldsymbol{V})$. Fix a reward function $\mathcal{R} \in \mathscr{R}$. First, the expect's performance $\mathbb{E}[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*]$ is obtainable from $P(\boldsymbol{V})$ by computing the arithmetic mean of $\mathcal{R}(\boldsymbol{Y})$ weighted by the marginal distribution $P(\boldsymbol{Y})$. If the NUC assumption (Def. 13) holds, the imitator's performance $\mathbb{E}_\pi[\mathcal{R}(\boldsymbol{Y}); \mathcal{M}^*]$ is computable from the observational distribution $P(\boldsymbol{V})$, following off-policy learning algorithms including IPW (Thm. 2) and DP (Thm. 3). The imitator could then formulate the minimax program in Eq. 474 from the observational distribution $P(\boldsymbol{V})$, the hypothesis reward class $\mathscr{R}$, and the policy space Π. Solving this optimization program leads to an imitating policy. We demonstrate in Example 65 the IRL strategy under NUC in an MAB environment.

On the other hand, however, when the input variables of the expert and the imitator's policies mismatch, and unobserved confounders generally exist, performing IRL with the standard off-policy evaluation does not necessarily lead to an imitating policy achieving the expert's performance. The following example illustrates the challenges of unobserved confounders for IRL methods.

Example 66 (Inverse RL fails without NUC) *Consider the alternative MAB environment $\mathcal{M}$ described in Eq. 462 where the front car's deceleration U is an unobserved confounder affecting both the human demonstrator's action X and the distance Y between the demonstrator and the front car; so the NUC does not hold in this environment.*

Evaluating the expected value of Y in this MAB environment $\mathcal{M}$ gives

$$\mathbb{E}[Y] = \mathbb{E}[X \oplus U] \tag{482}$$
$$= \mathbb{E}[\neg U \oplus U] \tag{483}$$

Computing the above equation gives the evaluation $\mathbb{E}[Y] = 1$. Applying the DP formula in Thm. 3 (or the IPW formula in Thm. 2) gives the following evaluation, for any policy $\pi(X)$,

$$\mathbb{E}_{\pi}[Y] = \mathbb{E}[Y \mid X = 0]\pi(X = 0) + \mathbb{E}[Y \mid X = 1]\pi(X = 1) \tag{484}$$

Among the above quantities, the conditional mean $\mathbb{E}[Y \mid X]$ is given by, for any x,

$$\mathbb{E}[Y \mid X = x] = \mathbb{E}[X \oplus U \mid X = x] \tag{485}$$
$$= \mathbb{E}[x \oplus \neg x \mid X = x] \tag{486}$$

The last step holds since values of X are given by $X \leftarrow \neg U$ in the MAB environment $\mathcal{M}$. Computing the above equation gives $\mathbb{E}[Y \mid X = x] = 1$ for $x = 0, 1$. Eq. 484 could be further written as

$$\mathbb{E}_{\pi}[Y] = \pi(X = 0) + \pi(X = 1) = 1 \tag{487}$$

Again, let $\mathcal{R}(Y) = \alpha Y$ be a linear reward function with a positive coefficient $\alpha > 0$. By substituting evaluations $\mathbb{E}[Y] = 1$ and $\mathbb{E}_{\pi}[Y] = 1$ into Eq. 476, we obtain the following minimax program:

$$\nu^* = \min_{\pi(X)} \max_{\alpha > 0} \alpha\mathbb{E}[Y] - \alpha\mathbb{E}_{\pi}[Y] \tag{488}$$
$$= \min_{\pi(X)} \max_{\alpha > 0} \alpha - \alpha \tag{489}$$
$$= 0 \tag{490}$$

This means that any fixed coefficient $\alpha > 0$, the imitator is able to achieve the expert's performance using any policy $\pi(x)$. To verify this conclusion, let an IRL policy $\pi_{\text{IRL}} : X \leftarrow 1$. Evaluating the expected reward $\mathcal{R}(Y)$ in submodel $\mathcal{M}_{\pi_{\text{IRL}}}$ implies

$$\mathbb{E}_{\pi_{\text{IRL}}}[\mathcal{R}(Y)] = \alpha\mathbb{E}_{X \leftarrow 1}[Y] \tag{491}$$
$$= \alpha\mathbb{E}[1 \oplus U] \tag{492}$$
$$= 0.5\alpha \tag{493}$$

The last step holds since U is uniformly drawn over the binary domain $\{0, 1\}$. This means that the IRL policy ($\mathbb{E}_{\pi_{\text{IRL}}}[\mathcal{R}(Y)] = 0.5\alpha$) fails to achieve the expert's performance $\mathbb{E}[\mathcal{R}(Y)] = \alpha$. ■

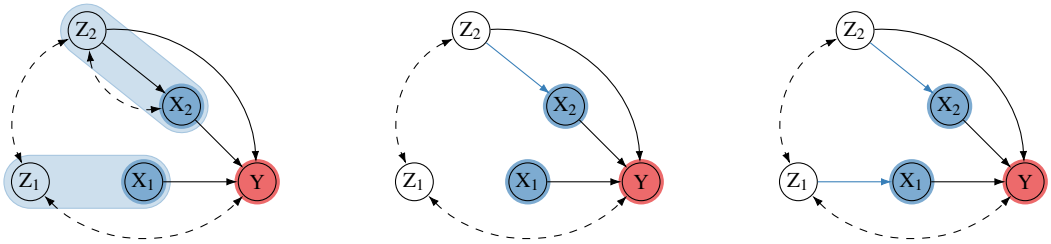


Figure 52: Causal diagrams where X represents an action (shaded blue) and Y represents a latent reward (shaded red). Input covariates of the policy space Π are shaded in light blue.

8.2.1 MINIMAL IMITATION BACKDOOR

We will next study causal IRL in more general settings where the NUC assumption does not hold, and there exist unobserved confounders in the demonstration data affecting both actions and other variables in the environment. Our algorithm relies on a refinement of the imitation backdoor condition (Def. 30), based on the concept of minimal d-separating sets.

Definition 32 (Minimal Imitation Backdoor) *Let $\mathcal{G}$ be a causal diagram and $\boldsymbol{X}, \boldsymbol{Y} \in \boldsymbol{V}$ be subsets of variables. An imitation admissible space Π over $\boldsymbol{X}$ is said to be* minimal *if there exists no proper subspace $\Pi' \subset \Pi$ satisfying the imitation backdoor w.r.t. $\boldsymbol{Y}$ in $\mathcal{G}$.* ■

In words, an imitation admissible space $\Pi = \{\langle X_i, \boldsymbol{S}_i\rangle\}_{X_i \in \boldsymbol{X}}$ is minimal if for every action $X_i \in \boldsymbol{X}$, $\boldsymbol{S}_i$ is a minimal d-separating set between action X_i and reward signals $\boldsymbol{Y}$ is the manipulated diagram $\mathcal{G}_{\underline{X_i},\pi_{i+1},\ldots,\pi_H}$; or states $\boldsymbol{S}_i = \emptyset$ whenever X_i is not an ancestor of $\boldsymbol{Y}$ in diagram $\mathcal{G}_{\pi_{i+1},\ldots,\pi_H}$.

Example 67 *Consider the causal diagram $\mathcal{G}$ described in Fig. 52a and a policy space $\Pi_1 = \{\langle X_1, \{Z_1\}\rangle, \langle X_2, \{Z_2\}\rangle\}$. For a policy $(\pi_1, \pi_2) \in \Pi$, the manipulated diagram $\mathcal{G}_{\pi_2}$ is shown in Fig. 52b. It is verifiable that Π_1 satisfies the imitation backdoor condition w.r.t. the outcome Y in $\mathcal{G}$ since the following independence relationships hold: $(X_1 \perp\!\!\!\perp Y \mid Z_1)$ in $\mathcal{G}_{\underline{X_1},\pi_2}$ and $(X_2 \perp\!\!\!\perp Y \mid Z_2)$ in $\mathcal{G}_{\underline{X_2}}$, respectively. However, the same space Π_1 is not minimal since $\{Z_1\}$ is not a minimal d-separating set and $(X_1 \perp\!\!\!\perp Y)$ holds in $\mathcal{G}^{(1)}_{\underline{X_1}}$. On the other hand, $\Pi_2 = \{\langle X_1, \emptyset\rangle, \langle X_2, \{Z_2\}\rangle\}$ is minimal imitation admissible since conditioning on the covariate set $\{Z_2\}$ d-separates the backdoor path $X_2 \leftarrow Z_2 \rightarrow Y$ in diagram $\mathcal{G}_{\underline{X_2}}$; removing node Z_2 opens the backdoor path.* ■

A key property of a minimal imitation admissible space Π is that for every policy $\pi \sim \Pi$, the interventional distribution $P_\pi(\boldsymbol{Y})$ is identifiable from the observational distribution $P(\boldsymbol{V})$, provided with the structural assumptions encoded in the causal diagram $\mathcal{G}$.

Theorem 21 *Let $\mathcal{G}$ be a causal diagram, Π be a policy space over actions $\boldsymbol{X}$, and $\boldsymbol{Y} \subseteq \boldsymbol{V}$ be a subset of variables. If there exists a subspace $\Pi' = \{\langle X_i, \boldsymbol{Z}_i\rangle\}_{i=1}^{H}$ contained in Π such that Π' is minimal imitation admissible w.r.t. $\boldsymbol{Y}$ in $\mathcal{G}$, then for every policy $\pi \in \Pi'$, the interventional distribution $P_\pi(\boldsymbol{Y})$ is computable from $P(\boldsymbol{V})$ and given by*

$$P_\pi(\boldsymbol{y}) = \sum_{\bar{\boldsymbol{x}}_H, \bar{\boldsymbol{z}}_H} P(\boldsymbol{y} \mid \bar{\boldsymbol{x}}_H, \bar{\boldsymbol{z}}_H) \prod_{i=1}^{H} P(\boldsymbol{z}_i \mid \bar{\boldsymbol{x}}_{i-1}, \bar{\boldsymbol{z}}_{i-1})\, \pi_i(x_i \mid \boldsymbol{z}_i)\,. \tag{494}$$

where $\bar{\boldsymbol{X}}_i = \{X_1, \ldots, X_i\}$ and $\bar{\boldsymbol{Z}}_i = \{\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_i\}$ are sequences of actions and input covariates up to the decision horizon $i = 1, \ldots, H$. ■

However, the same identifiability result does not generally hold for policies in a non-minimal imitation admissible space. The following example demonstrates such an instance.

Example 68 *Consider the causal diagram $\mathcal{G}$ of Fig. 52a again. Let us focus on the minimal imitation admissible space*

$$\Pi_2 = \{\langle X_1, \emptyset\rangle, \langle X_2, \{Z_2\}\rangle\} \tag{495}$$

For ever policy $\pi \in \Pi_2$, the post-interventional diagram $\mathcal{G}_\pi$ is shown in Fig. 52b. By applying Thm. 21 we obtain

$$P_\pi(y) = \sum_{x_1, x_2, z_2} P(y \mid x_1, x_2, z_2)\, P(z_2 \mid x_1)\pi_2(x_2 \mid z_2)\pi_1(x_1) \tag{496}$$

On the other hand, the same identification result in Thm. 21 does not necessarily hold for a non-minimal imitation admissible space.

More specifically, consider a policy space

$$\Pi_1 = \{\langle X_1, \{Z_1\}\rangle, \langle X_2, \{Z_2\}\rangle\} \tag{497}$$

For ever policy $\pi \in \Pi_1$, the post-interventional diagram $\mathcal{G}_\pi$ is shown in Fig. 52c. As discussed previously (Example 67), Π_1 is not minimal. This means that, for any policy $\pi \in \Pi_1$, the interventional distribution $P_\pi(Y)$ is not computable from the identification formula in Eq. 494. More generally, $P_\pi(Y)$ for policies $\pi \in \Pi_1$ is not identifiable from the observational distribution $P(\boldsymbol{V})$ in diagram $\mathcal{G}$. Following the decomposition in Eq. 212, $P_\pi(Y)$ can be written as,

$$P_\pi(y) = \sum_{x_1, x_2, z_1, z_2} P_{x_1, x_2}(y, z_1, z_2)\, \pi_1(x_1 \mid z_1)\pi_2(x_2 \mid z_2) \tag{498}$$

Prop. 3 implies that $P_\pi(Y)$ is identifiable if and only if the interventional distribution $P_{x_1,x_2}(Y, Z_1, Z_2)$ is identifiable in the causal diagram $\mathcal{G}$. However, such quantity $P_{x_1,x_2}(Y, Z_1, Z_2)$ is not identifiable due to the presence of the bi-directed path $X_2 \dashleftarrow\dashrightarrow Z_2 \dashleftarrow\dashrightarrow Z_1 \dashleftarrow\dashrightarrow Y$ (Tian, 2002, Thm. 16). Indeed, the non-identifiability of the effects of policies $\pi \in \Pi_1$ in the causal diagram $\mathcal{G}$ described in Fig. 52a has been shown in (Tian, 2008; Correa and Bareinboim, 2019). ■

The concept of minimal imitation backdoor in Def. 32 and the identification result in Thm. 21 provide a natural algorithm for performing causal IRL when unobserved confounders generally exist. Instead of searching for imitating policies in the policy space Π, the agent will focus on a minimal imitation admissible subspace $\Pi' \subseteq \Pi$. Specifically, as discussed previously in Sec. 8.1, there exist efficient algorithms finding admissible policy subspaces satisfying the imitation backdoor. Once such an admissible subspace is found, one could obtain a minimal imitation admissible subspace by iteratively removing input state variables from $\boldsymbol{S}_i$ for every action X_i until the imitation backdoor does not hold. This procedure could be done in polynomial steps with regard to the total number of states $\boldsymbol{S}$ and actions $\boldsymbol{X}$ variables.

8.2.2 IMITATION VIA INVERSE RL

Once a minimal imitation admissible subspace $\Pi' \subseteq \Pi$ is obtained, one could obtain an imitating policy by solving the minimax program in Eq. 474 with the policy space Π substituted with Π'. By expanding values of $\boldsymbol{Y}$, this optimization program could be written as,

$$\nu^* = \min_{\pi \in \Pi'} \max_{\mathcal{R} \in \mathscr{R}} \sum_{\boldsymbol{y}} \mathcal{R}(\boldsymbol{y}) (\underbrace{P(\boldsymbol{y})}_{\text{expert's occupancy}} - \underbrace{P_\pi(\boldsymbol{y})}_{\text{imitator's occupancy}}) \tag{499}$$

Among quantities in the above equation, the first term is the expert's occupancy measures over domains of signals $\boldsymbol{Y}$, which is a marginal observational distribution $P(\boldsymbol{Y})$. The second term is the expert's occupancy measures over domains of $\boldsymbol{Y}$, which is an interventional distribution $P_\pi(\boldsymbol{Y})$. Since Π' is a minimal subspace satisfying the imitation backdoor criterion, applying Thm. 21 permits one to compute $P_\pi(\boldsymbol{Y})$ from the observational distribution $P(\boldsymbol{V})$ and the policy π.[65]

Provided with some common choices of the hypothesis class $\mathscr{R}$, the minimax program in Eq. 499 is solvable using some state-of-art IRL algorithms. Due to this reason, we consistently refer to Eq. 499 as the *canonical IRL program*. To make this argument more precise, we will demonstrate this reduction procedure with the multiplicative-weights algorithm (MWAL) (Syed and Schapire, 2008) and the generative adversarial imitation learning (GAIL) (Ho and Ermon, 2016).

Causal MWAL (Abbeel and Ng, 2004; Syed and Schapire, 2008) study IRL in Markov decision processes where the reward function $\mathcal{R}(\boldsymbol{y})$ is a linear combination of k-length *feature expectations* vectors $\boldsymbol{\phi}(\boldsymbol{y})$. Particularly, let $\mathcal{R}(\boldsymbol{y}) = \boldsymbol{w} \cdot \boldsymbol{\phi}(\boldsymbol{y})$ for a coefficient vector $\boldsymbol{w}$ in a convex set

$$\mathbb{P}^k = \left\{ \boldsymbol{w} \in \mathbb{R}^k \mid \|\boldsymbol{w}\|_1 = 1 \text{ and } \boldsymbol{w} \succeq \boldsymbol{0} \right\}. \tag{500}$$

Let $\boldsymbol{\phi}^{(i)}$ be the i-th component of feature vector $\boldsymbol{\phi}$ and let deterministic policies with space Π be ordered by $\pi^{(1)}, \ldots, \pi^{(n)}$. The canonical program in Eq. 499 is reducible to a two-person zero-sum matrix game under linearity.

Proposition 6 *For a hypothesis class $\mathscr{R} = \{\mathcal{R} = \boldsymbol{w} \cdot \boldsymbol{\phi} \mid \boldsymbol{w} \in \mathbb{P}^k\}$, the solution ν^* of the canonical program in Eq. 499 is obtainable by solving the following minimax problem*

$$\nu^* = \min_{\pi \in \Pi'} \max_{\boldsymbol{w} \in \mathbb{P}^k} \boldsymbol{w}^\top \boldsymbol{G} \pi, \tag{501}$$

where $\boldsymbol{G}$ is a $k \times n$ matrix given by $\boldsymbol{G}(i,j) = \sum_{\boldsymbol{y}} \boldsymbol{\phi}^{(i)}(\boldsymbol{y}) \left(P(\boldsymbol{y}) - P_{\pi^{(j)}}(\boldsymbol{y})\right)$. ■

There exist effective multiplicative weights algorithms for solving the matrix game in Eq. 501, including MW (Freund and Schapire, 1999) and MWAL (Syed and Schapire, 2008).

Causal GAIL (Ho and Ermon, 2016) introduces the GAIL algorithm for learning an imitating policy in Markov decision processes with a general family of non-linear reward functions. In particular, $\mathcal{R}(\boldsymbol{y})$ takes values in the real space $\mathbb{R}$, i.e., $\mathcal{R} \in \mathbb{R}^{\boldsymbol{Y}}$ where $\mathbb{R}^{\boldsymbol{Y}} = \{r : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}\}$. The complexity of reward function $\mathcal{R}$ is penalized by a convex regularization function $\psi(\mathcal{R})$, i.e.,

$$\nu^* = \min_{\pi \in \Pi'} \max_{\mathcal{R} \in \mathbb{R}^{\boldsymbol{Y}}} \sum_{\boldsymbol{y}} \mathcal{R}(\boldsymbol{y}) \left(P(\boldsymbol{y}) - P_\pi(\boldsymbol{y})\right) - \psi(\mathcal{R}) \tag{502}$$

65. More generally, the imitator could search over all policies $\pi \in \Pi$ such that the imitator's occupancy measure $P_\pi(\boldsymbol{Y})$ induced by $\text{do}(\pi)$ is identifiable in diagram $\mathcal{G}$. This imitation approach has been studied in (Ruan et al., 2023).

Henceforth, we will consistently refer to Eq. 502 as the *penalized canonical program* of causal IRL. It is often preferable to solve its conjugate form. Formally,

Proposition 7 *For a hypothesis class $\mathscr{R} = \{\mathcal{R} : \mathscr{D}(\boldsymbol{Y}) \mapsto \mathbb{R}\}$ regularized by ψ, the solution ν^* of the penalized canonical program in Eq. 502 is obtainable by solving the following problem*

$$\nu^* = \min_{\pi \in \Pi'} \psi^* \left(P - P_\pi\right) \tag{503}$$

where ψ^* be a conjugate function of ψ and is given by $\psi^* = \max_{\mathcal{R} \in \mathbb{R}^{\boldsymbol{Y}}} a^\top \mathcal{R} - \psi(\mathcal{R})$. ■

Eq. 503 seeks a policy π which minimizes the divergence of joint probabilities over reward signals $\boldsymbol{Y}$ between the imitator and the expert, as measured by the function ψ^*. When we utilize a regularizer $\psi(r)$ similar to (Ho and Ermon, 2016, Eq. 13), the convex conjugate function ψ^* in Eq. 503 is further written as:

$$\min_{\pi \in \Pi'} \psi^* \left(P - P_\pi\right) = \min_{\pi \in \Pi'} \max_{D \in (0,1)^{\boldsymbol{Y}}} E\left[\log(D(\boldsymbol{Y}))\right] + \mathbb{E}_\pi\left[\log(1 - D(\boldsymbol{Y}))\right], \tag{504}$$

where function $D \in \mathscr{D}(\boldsymbol{Y}) \mapsto (0, 1)$ is a discriminator classifier (e.g, a neural network). The above equation draws the connection between causal imitation learning and the computational framework of generative adversarial networks (Goodfellow et al., 2014), which could be viewed as two neural networks competing against each other in a zero-sum game. When the discriminator D cannot distinguish the occupancy measure generated by the policy π from the expert, then π has successfully matched the expert's performance. Solving the minimax program of Eq. 504 requires finding a saddle point (π, D). This could be done by iteratively optimizing policy parameters π and discriminator D following the implementation procedure of GAIL algorithm (Ho and Ermon, 2016).

Experiment 13 *We demonstrate our causal imitation framework on an SCM $\mathcal{M}^*$ compatible with the causal diagram in Fig. 52a. Particularly,*

$$\mathcal{M}^* = \langle \boldsymbol{U} = \{U_1, U_2, U_3, U_4\}, \boldsymbol{L} = \emptyset, \boldsymbol{V} = \{X_1, X_2, Z_1, X_2, Y\}, \mathscr{F}, P(\boldsymbol{U})\rangle \tag{505}$$

where structural functions $\mathscr{F}$ is defined as

$$\mathscr{F} = \begin{cases} Z_1 \leftarrow U_1 \oplus U_3, \\ X_1 \sim \texttt{Bern}(0.68) \\ Z_2 \leftarrow U_1 \oplus U_2 \oplus U_4, \\ X_2 \leftarrow U_2 \oplus Z_2 \\ Y \leftarrow (X_1, X_2, Z_1, Z_2, U_3) \end{cases} \tag{506}$$

Among quantities in the above equation, reward signal $Y = (Y_1, \dots Y_5)$ is a feature vector containing 5 elements; the exogenous distribution $P(U_1, U_2, U_3, U_4)$ is defined such that U_i, $i = 1, \dots, 4$ are independent variables given by

$$U_1 \sim \texttt{Bern}(0.8), \qquad U_2 \sim \texttt{Bern}(0.8), \qquad U_3 \sim \texttt{Bern}(0.2) \qquad U_4 \sim \texttt{Bern}(0.1) \tag{507}$$

The agent's goal is to optimize a CDM $\langle \mathcal{M}^*, \Pi, \mathcal{R}\rangle$ where environment $\mathcal{M}^*$ is defined in Eq. 505; the policy space $\Pi = \{\langle X_1, \{Z_1\}\rangle, \langle X_2, \{Z_1, X_1, Z_2\}\rangle\}$; and the reward function $\mathcal{R}(Y) = \oplus_{i=1}^5 Y_i$.

We will then apply different imitation strategies to learn an imitating policy in space Π without detailed parametrization of the reward function $\mathcal{R}(Y)$. These imitation algorithms are

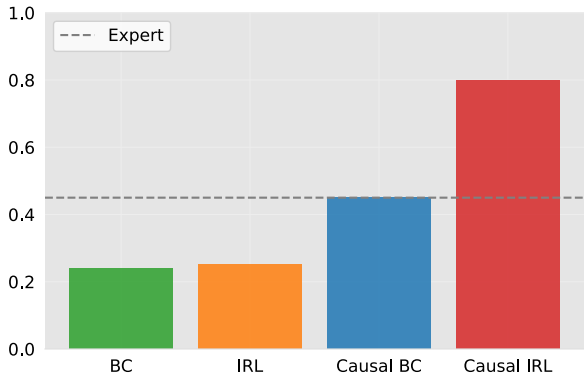


Figure 53: Simulation results evaluating causal IRL when imitation backdoor condition holds.

- *Standard BC algorithm utilizes all observed states $\boldsymbol{S}_i$ for every action $X_i \in \boldsymbol{X}$.*
- *Standard IRL algorithm utilizes all observed states $\boldsymbol{S}_i$ for every action $X_i \in \boldsymbol{X}$. We will apply GAIL algorithm (Syed and Schapire, 2008) when $\mathcal{R}(Y)$ is non-linear;*
- *Causal-BC algorithm, described in Thm. 20, selects a set of covariates $\boldsymbol{Z}_i$ for every action $X_i \in \boldsymbol{X}$ following the imitation backdoor criterion (Def. 30). It then learns an imitating policy π with subspace $\Pi' = \{\langle X_i, \boldsymbol{Z}_i \rangle\}_{X_i \in \boldsymbol{X}}$ using standard BC algorithms.*
- *Our proposed Causal-IRL algorithm first finds a minimal imitation admissible subspace Π' (Thm. 21) and then obtains an imitating policy by solving the canonical program in Eq. 499. We will use Causal GAIL algorithm since $\mathcal{R}(Y)$ is non-linear. Reward augmentation (RA) is performed to incorporate the parametric knowledge that $\mathcal{R}(Y)$ is a monotone function concerning values of Y_1, Y_2 (Li et al., 2017). This is done by adding an additional regularization function in Eq. 504 to encourage assigning higher values of features Y_1, Y_2.*

Simulation results are shown in Fig. 53. The analysis reveals that Causal-IRL consistently outperforms the expert's policy and other imitation strategies by exploiting additional parametric knowledge about the reward function; Causal-BC obtains a policy that mimics the expert's performance. As expected, BC and IRL failed to obtain a policy that matches the expert's performance. ■

This section investigates imitation learning in the semantics of structural causal models. The goal is to find an imitating policy that can perform at least as well as the expert behaviors from combinations of demonstration data and qualitative knowledge about the data-generating process represented as a causal diagram. First, we provided a novel graphical criterion that is sufficient for determining the feasibility of learning an imitating policy that mimics the expert's performance. When such a condition holds, one could obtain an effective imitating learning using standard behavioral cloning. We also investigate imitation learning via inverse reinforcement learning (IRL), provided with additional quantitative knowledge about the reward function. We provide a graphical criterion based on the sequential backdoor, which allows one to obtain an imitating policy by solving a canonical optimization equation of causal IRL. Such a canonical formulation addresses the challenge of the presence of unobserved confounders (UCs) and is solvable by leveraging standard IRL algorithms.

9. Conclusions

The current generation of AI agents capable of optimal decision-making builds on the theoretical framework of reinforcement learning (RL). Most of these RL systems do not explicitly represent the underlying causal models or engage in causal reasoning. On the other hand, there is a growing recognition across many fields and sciences that effective decision-making relies on an understanding of the causal mechanisms in the environment. For example, an intelligent robot needs to grasp the cause-and-effect relationships within its surroundings to plan its actions effectively; a physician must understand the effects of available medications to devise a suitable treatment strategy for her patients; an economist, too, needs to envision the relationship between skill sets and the future job market in order to create an effective educational policy. These scenarios illustrate how decision-making across various sectors of society depends on understanding complex, dynamic, and often unobserved causal mechanisms. Although there have been some attempts to integrate causal knowledge into RL tasks, a systematic approach and a cohesive foundation are still lacking.

To address this challenge, we combine the capabilities of RL agents with Pearl's Structural Causal Models (SCMs) theory to encode causal knowledge and perform counterfactual reasoning. This marriage leads to an algorithmic and theoretical framework for robust decision-making under uncertainties, which is part of an emerging branch of research called *Causal Reinforcement Learning* (CRL). Building on this framework, we are able to bring improvement to RL algorithms in some key aspects. First, The CRL framework enables us to relax certain key assumptions regarding the causal mechanisms that generate the observed data. We have developed innovative algorithms that are resilient to unobserved confounding bias in offline settings, which include off-policy learning and imitation learning. Additionally, we proposed more efficient online learning algorithms that effectively identify optimal policies while achieving near-optimal regret. These advancements leverage causal conclusions drawn from biased offline data.

The final important distinction presented in this manuscript is the difference between the regimes in which AI agents operate to interact with their environment. Specifically, supervised learning agents identify patterns from data gathered through passive observation, while reinforcement learning (RL) agents actively engage with the system and modify their policies based on the responses they receive. By generalizing these interaction regimes, we open up new learning possibilities that have not been explored in the existing literature. The problem of where to intervene allows us to design agents to achieve better performance by combining both passive observation and active intervention. Counterfactual randomization generalizes the classic Fisherian randomized experience, enriching agents with capabilities of counterfactual reasoning, which we believe is critical to design the next generation of AI systems.

Acknowledgements

This research was supported in part by the NSF, ONR, AFOSR, DARPA, DoE, Amazon, JP Morgan, and The Alfred P. Sloan Foundation.

References

Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM, 2004.

Raj Agrawal, Chandler Squires, Karren D. Yang, Karthikeyan Shanmugam, and Caroline Uhler. Abcd-strategy: Budgeted experimental design for targeted causal structure discovery. In *AISTATS*, pages 3400–3409, 2019. URL http://proceedings.mlr.press/v89/agrawal19b.html.

Shipra Agrawal and Navin Goyal. Analysis of thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, pages 39–1, 2012.

T. Anand, A. Ribeiro, J. Tian, and E. Bareinboim. Effect identification in causal diagrams with clustered variables. Technical Report R-77, Causal Artificial Intelligence Lab, Columbia University. https://causalai.net/r77.pdf, Jun 2021.

Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

Karl Johan Åström. Optimal control of markov processes with incomplete state information. *Journal of mathematical analysis and applications*, 10(1):174–205, 1965.

Jean-Yves Audibert and Sébastien Bubeck. Minimax policies for adversarial and stochastic bandits. In *COLT*, pages 217–226, 2009.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2/3):235–256, 2002a.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2):235–256, 2002b.

Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multi-armed bandit problem. *SIAM journal on computing*, 32(1):48–77, 2002c.

Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. In *Advances in neural information processing systems*, pages 89–96, 2009.

Mohammad Gheshlaghi Azar, Ian Osband, and Rémi Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, pages 263–272. PMLR, 2017.

Philip J Ball, Laura Smith, Ilya Kostrikov, and Sergey Levine. Efficient online reinforcement learning with offline data. *arXiv preprint arXiv:2302.02948*, 2023.

E. Bareinboim and J. Pearl. Causal inference by surrogate experiments: z-identifiability. In Nando de Freitas and Kevin Murphy, editors, *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 113–120, Corvallis, OR, 2012a. AUAI Press.

Elias Bareinboim and Judea Pearl. Causal inference by surrogate experiments: z-identifiability. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 113–120, 2012b.

Elias Bareinboim and Judea Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27):7345–7352, 2016.

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

Elias Bareinboim, JD Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy and the foundations of causal inference. *ACM Special Volume in Honor of Judea Pearl (provisional title)*, 2020.

John A. Bargh and Tanya L. Chartrand. The unbearable automaticity of being. *American Psychologist*, 54(7):462–479, 1999.

Richard Bellman. *Dynamic Programming*. Princeton University Press, Princeton, NJ, 1957.

Richard Bellman. Dynamic programming. *Science*, 153(3731):34–37, 1966.

Andrew Bennett, Nathan Kallus, Lihong Li, and Ali Mousavi. Off-policy evaluation in infinite-horizon reinforcement learning with latent confounders. In *International Conference on Artificial Intelligence and Statistics*, pages 1999–2007. PMLR, 2021.

Christopher Berner, Greg Brockman, Brooke Chan, Vicki Cheung, Przemyslaw Debiak, Christy Dennison, David Farhi, Quirin Fischer, Shariq Hashme, Chris Hesse, et al. Dota 2 with large scale deep reinforcement learning. *arXiv preprint arXiv:1912.06680*, 2019.

Donald A Berry and Bert Fristedt. Bandit problems: sequential allocation of experiments (monographs on statistics and applied probability). *London: Chapman and Hall*, 5(71-87):7–7, 1985.

Dimitri P Bertsekas. *Dynamic programming and optimal control*, volume 1. Athena scientific Belmont, MA, 3 edition, 2005.

Dimitri P Bertsekas and John N Tsitsiklis. Neuro-dynamic programming: an overview. In *Proceedings of 1995 34th IEEE conference on decision and control*, volume 1, pages 560–564. IEEE, 1995.

Aude Billard, Sylvain Calinon, Ruediger Dillmann, and Stefan Schaal. Survey: Robot programming by demonstration. *Handbook of robotics*, 59(BOOK_CHAP), 2008.

Bibhas Chakraborty and Erica E Moodie. Statistical methods for dynamic treatment regimes. *Springer-Verlag. doi*, 10:978–1, 2013.

Bibhas Chakraborty and Susan A Murphy. Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447, 2014.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, pages 2249–2257, 2011.

Eunah Cho. *Robust Causal Inference Methods for Using Randomized Clinical Trial and Observational Study*. North Carolina State University, 2022.

Bénédicte Colnet, Imke Mayer, Guanhua Chen, Awa Dieng, Ruohong Li, Gaël Varoquaux, Jean-Philippe Vert, Julie Josse, and Shu Yang. Causal inference methods for combining randomized trials and observational studies: a review. *arXiv preprint arXiv:2011.08047*, 2020.

Thomas H Cormen, Charles E Leiserson, Ronald L Rivest, and Clifford Stein. *Introduction to algorithms*. MIT press, 2022.

J. Cornfield. A method of estimating comparative rates from clinical data; applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute*, 11:1269–1275, 1951.

Jerome Cornfield, William Haenszel, E Cuyler Hammond, Abraham M Lilienfeld, Michael B Shimkin, and Ernst L Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions. *Journal of the National Cancer institute*, 22(1):173–203, 1959.

Juan Correa and Elias Bareinboim. A calculus for stochastic interventions: Causal effect identification and surrogate experiments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10093–10100, 2020a.

Juan Correa and Elias Bareinboim. General transportability of soft interventions: Completeness results. *Advances in Neural Information Processing Systems*, 33:10902–10912, 2020b.

Juan D Correa and Elias Bareinboim. From statistical transportability to estimating the effect of stochastic interventions. In *IJCAI*, 2019.

A.P. Dawid. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70:161–189, 2002.

Pim de Haan, Dinesh Jayaraman, and Sergey Levine. Causal confusion in imitation learning. In *Advances in Neural Information Processing Systems*, pages 11693–11704, 2019.

F d'Epenoux. A probabilistic production and inventory problem. *Management Science*, 10(1): 98–108, 1963.

Vanessa Didelez, A. Philip Dawid, and Sara Geneletti. Direct and indirect effects of sequential treatments. In *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence*, page 138–146. AUAI Press, 2006.

Thomas G Dietterich. Hierarchical reinforcement learning with the maxq value function decomposition. *Journal of artificial intelligence research*, 13:227–303, 2000.

Ap Dijksterhuis and Loran F. Nordgren. A theory of unconscious thought. *Perspectives on Psychological Science*, 1(2):95–109, 2006.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 2005.

Jalal Etesami and Philipp Geiger. Causal transfer for imitation learning and decision making under sensor-shift. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, New York, NY, 2020. AAAI Press.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. Building watson: An overview of the deepqa project. *AI magazine*, 31(3):59–79, 2010.

Jerzy Filar and Koos Vrieze. *Competitive Markov decision processes*. Springer Science & Business Media, 2012.

R.A. Fisher. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain*, 33:503–513, 1926.

R.A. Fisher. *The Design of Experiments*. Oliver and Boyd, Edinburgh, 1935.

Katherine M. Flegal, Barry I. Graubard, and David F. Williamson. Excess deaths associated with underweight, overweight, and obesity. 293(15):1861–1867, 2005.

Andrew Forney and Elias Bareinboim. Counterfactual randomization: Rescuing experimental studies from obscured confounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2454–2461, 2019.

Andrew Forney, Judea Pearl, and Elias Bareinboim. Counterfactual data-fusion for online reinforcement learners. In *International Conference on Machine Learning*, pages 1156–1164. PMLR, 2017.

Yoav Freund and Robert E Schapire. Adaptive game playing using multiplicative weights. *Games and Economic Behavior*, 29(1-2):79–103, 1999.

Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

John C Gittins. Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 148–177, 1979.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Carlos Guestrin, Daphne Koller, and Ronald Parr. Multiagent planning with factored mdps. *Advances in neural information processing systems*, 14, 2001.

Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored mdps. *Journal of Artificial Intelligence Research*, 19:399–468, 2003.

T. Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11: 1–12, 1943.

Eric A Hansen. Solving pomdps by searching in policy space. In *Proceedings of the Fourteenth conference on Uncertainty in artificial intelligence*, pages 211–219, 1998.

Milos Hauskrecht. Value-function approximations for partially observable markov decision processes. *Journal of artificial intelligence research*, 13:33–94, 2000.

L. Hayduk, G. Cummings, R. Stratkotter, M. Nimmo, K. Grygoryev, D. Dosman, M. Gillespie, and H. Pazderka-Robinson. Pearls d-separation: One more step into causal thinking. *Structural Equation Modeling*, 10(2):289–311, 2003.

James J Heckman. Randomization and social policy evaluation. *Evaluating welfare and training programs*, 1:201–30, 1992.

James J. Heckman. Skill formation and the economics of investing in disadvantaged children. *Science*, 312(5782):1900–1902, 2006.

Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29, 2016.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Y. Huang and M. Valtorta. Identifiability in causal bayesian networks: A sound and complete algorithm. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence (AAAI 2006)*, pages 1149–1156. AAAI Press, Menlo Park, CA, 2006a.

Y. Huang and M. Valtorta. Pearl's calculus of intervention is complete. In R. Dechter and T.S. Richardson, editors, *Proceedings of the Twenty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2006)*, pages 217–224. AUAI Press, Corvallis, OR, 2006b.

Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)*, 50(2):1–35, 2017.

Guido W Imbens. Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and statistics*, 86(1):4–29, 2004.

Tommi Jaakkola, Satinder Singh, and Michael Jordan. Reinforcement learning algorithm for partially observable markov decision problems. *Advances in neural information processing systems*, 7, 1994.

Tommi Jaakkola, Satinder P Singh, and Michael I Jordan. Reinforcement learning algorithm for partially observable markov decision problems. In *Advances in neural information processing systems*, pages 345–352, 1995.

A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9551–9561, Vancouver, Canada, Jun 2020. Curran Associates, Inc.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600, 2010.

Y. Jung, I. Diaz, J. Tian, and E. Bareinboim. Estimating causal effects identifiable from combination of observations and experiments. Technical Report R-97, Causal Artificial Intelligence Lab, Columbia University, May 2023a.

Y. Jung, J. Tian, and E. Bareinboim. Estimating joint treatment effects by combining multiple experiments. In *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, Hawaii, Jul 2023b.

Lodewijk Cornelis Maria Kallenberg. Linear programming and finite markovian control problems. *MC Tracts*, 1983.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 9289–9299, 2018.

Nathan Kallus and Angela Zhou. Confounding-robust policy evaluation in infinite-horizon reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 22293–22304. Curran Associates, Inc., 2020.

Michael Kearns, Yishay Mansour, and Andrew Ng. Approximate planning in large pomdps via reusable trajectories. *Advances in Neural Information Processing Systems*, 12, 1999.

M. Kocaoglu, A. Jaber, K. Shanmugam, and E. Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 14346–14356, Vancouver, Canada, 2019. Curran Associates, Inc.

Murat Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental design for learning causal graphs with latent variables. In *Advances in Neural Information Processing Systems*, pages 7018–7028, 2017.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Daphne Koller and Brian Milch. Multi-agent influence diagrams for representing and solving games. *Games and economic behavior*, 45(1):181–221, 2003.

Sid Kouider, Vincent de Gardelle, Jérôme Sackur, and Emmanuel Dupoux. How rich is consciousness? the partial awareness hypothesis. *Trends in Cognitive Sciences*, 14(7):301–307, 2010.

Robert Krajewski, Julian Bock, Laurent Kloeker, and Lutz Eckstein. The highd dataset: A drone dataset of naturalistic vehicle trajectories on german highways for validation of highly automated driving systems. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2118–2125, 2018. doi: 10.1109/ITSC.2018.8569552.

Tejas D Kulkarni, Karthik Narasimhan, Ardavan Saeedi, and Josh Tenenbaum. Hierarchical deep reinforcement learning: Integrating temporal abstraction and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

Aviral Kumar, Joey Hong, Anikait Singh, and Sergey Levine. When should we prefer offline reinforcement learning over behavioral cloning? *arXiv preprint arXiv:2204.05618*, 2022.

Daniel Kumor, Junzhe Zhang, and Elias Bareinboim. Sequential causal imitation learning with unobserved confounders. *Advances in Neural Information Processing Systems*, 2021.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985a.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22, 1985b.

Sascha Lange, Thomas Gabel, and Martin Riedmiller. Batch reinforcement learning. In *Reinforcement Learning*, pages 45–73. Springer, 2012.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In J.C. Platt, D. Koller, Y. Singer, and S.T. Roweis, editors, *Advances in Neural Information Processing Systems 20*, pages 817–824. Curran Associates, Inc., 2008a.

John Langford and Tong Zhang. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in neural information processing systems*, pages 817–824, 2008b.

Jean B Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Steffen L Lauritzen and Dennis Nilsson. Representing and solving decision problems with limited information. *Management Science*, 47(9):1235–1251, 2001.

Alessandro Lazaric. Transfer in reinforcement learning: a framework and a survey. In *Reinforcement Learning*, pages 143–173. Springer, 2012.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? Technical Report R-36, Purdue AI Lab, Department of Computer Science, Purdue University, 2018a.

Sanghack Lee and Elias Bareinboim. Structural causal bandits: where to intervene? In *Advances in Neural Information Processing Systems*, pages 2568–2578, 2018b.

Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. Technical Report R-40, Purdue AI Lab, Department of Computer Science, Purdue University, 2019a.

Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 4164–4172, 2019b.

Sanghack Lee and Elias Bareinboim. Characterizing optimal mixed policies: Where to intervene and what to observe. *Advances in Neural Information Processing Systems*, 33, 2020.

Sanghack Lee, Juan David Correa, and Elias Bareinboim. General identifiability with arbitrary surrogate experiments. In *In Proceedings of the 35th Conference on Uncertainty in Artificial Intelligence*, 2019.

Seunghyun Lee, Younggyo Seo, Kimin Lee, Pieter Abbeel, and Jinwoo Shin. Offline-to-online reinforcement learning via balanced replay and pessimistic q-ensemble. In *Conference on Robot Learning*, pages 1702–1712. PMLR, 2022.

D. Lewis. *Counterfactuals*. Harvard University Press, Cambridge, MA, 1973.

Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

Lihong Li, Wei Chu, John Langford, and Xuanhui Wang. Unbiased offline evaluation of contextual-bandit-based news article recommendation algorithms. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 297–306. ACM, 2011.

Lihong Li, Remi Munos, and Csaba Szepesvári. On minimax optimal offline policy evaluation. *arXiv preprint arXiv:1409.3653*, 2014.

Yunzhu Li, Jiaming Song, and Stefano Ermon. Infogail: Interpretable imitation learning from visual demonstrations. *Advances in Neural Information Processing Systems*, 30, 2017.

Benjamin Libet, Curtis A Gleason, Elwood W Wright, and Dennis K Pearl. Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). In *Neurophysiology of consciousness*, pages 249–268. Springer, 1993.

Xi Lin and Robin J Evans. Many data: Combine experimental and observational data through a power likelihood. *arXiv preprint arXiv:2304.02339*, 2023.

Michael L Littman, Anthony R Cassandra, and Leslie Pack Kaelbling. Learning policies for partially observable environments: Scaling up. In *Machine Learning Proceedings 1995*, pages 362–370. Elsevier, 1995.

Qiang Liu and Alexander Ihler. Belief propagation for structured decision making. In *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 523–532. AUAI Press, 2012.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. *Advances in neural information processing systems*, 31, 2018.

William S Lovejoy. A survey of algorithmic methods for partially observed markov decision processes. *Annals of Operations Research*, 28(1):47–65, 1991.

Jeffrey Mahler and Ken Goldberg. Learning deep policies for robot bin picking by simulating robust grasping sequences. In *Conference on robot learning*, pages 515–524, 2017.

C.F. Manski. Nonparametric bounds on treatment effects. *American Economic Review, Papers and Proceedings*, 80:319–323, 1990.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Joris M. Mooij, Sara Magliacane, and Tom Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 21(99):1–108, 2020.

Thomas J Moore, Hanzhe Zhang, Gerard Anderson, and G Caleb Alexander. Estimated costs of pivotal trials for novel therapeutic agents approved by the us food and drug administration, 2015-2016. *JAMA internal medicine*, 178(11):1451–1457, 2018.

Stanley A Mulaik. *Linear causal modeling with structural equations*. Chapman and Hall/CRC, 2009.

Urs Muller, Jan Ben, Eric Cosatto, Beat Flepp, and Yann L Cun. Off-road obstacle avoidance through end-to-end learning. In *Advances in neural information processing systems*, pages 739–746, 2006.

Katharina Mülling, Jens Kober, Oliver Kroemer, and Jan Peters. Learning to select and generalize striking movements in robot table tennis. *The International Journal of Robotics Research*, 32(3): 263–279, 2013.

S. A. Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

S A Murphy, M J van der Laan, J M Robins, and Conduct Problems Prevention Research Group. Marginal Mean Models for Dynamic Regimes. *Journal of the American Statistical Association*, 96(456):1410–1423, December 2001a.

Susan A Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10):1455–1481, 2005a.

Susan A Murphy. A generalization error for Q-learning. *Journal of Machine Learning Research*, 6 (Jul):1073–1097, 2005b.

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

Inbal Nahum-Shani, Min Qian, Daniel Almirall, William E Pelham, Beth Gnagy, Gregory A Fabiano, James G Waxmonsky, Jihnhee Yu, and Susan A Murphy. Experimental design and primary data analysis methods for comparing adaptive interventions. *Psychological methods*, 17(4):457, 2012.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *Advances in Neural Information Processing Systems*, 33:18819–18831, 2020.

J. Neyman. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, 5(4):465–480, 1923.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pages 278–287, 1999.

Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In *Icml*, volume 1, pages 663–670, 2000.

Dirk Ormoneit and Aunak Sen. Kernel-based reinforcement learning. *Machine learning*, 49(2-3): 161, 2002.

Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, Jan Peters, et al. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 7 (1-2):1–179, 2018.

Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored mdps. In *Advances in Neural Information Processing Systems*, pages 604–612, 2014.

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

J. Pearl. Aspects of graphical models connected with causality. In *Proceedings of the 49th Session of the International Statistical Institute*, pages 391–401, Tome LV, Book 1, Florence, Italy, 1993.

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

J. Pearl and J.M. Robins. Probabilistic evaluation of sequential plans from causal models with hidden variables. In P. Besnard and S. Hanks, editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (UAI 1995)*, pages 444–453. Morgan Kaufmann, San Francisco, 1995.

Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. Basic Books, 2018.

K. Pearson. *Grammar of Science,* 3rd ed. A. and C. Black Publishers, London, 1911.

M.L. Petersen, S.E. Sinisi, and M.J. van der Laan. Estimation of direct causal effects. *Epidemiology*, 17(3):276–284, 2006.

D. Plecko and E. Bareinboim. Causal fairness analysis. Technical Report R-90, Causal Artificial Intelligence Lab, Columbia University, Jul 2022.

Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In *Advances in neural information processing systems*, pages 305–313, 1989.

Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., 1994.

Jette Randløv and Preben Alstrøm. Learning to drive a bicycle using reinforcement learning and shaping. In *ICML*, volume 98, pages 463–471. Citeseer, 1998.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Soc.*, 58 (5):527–535, 09 1952.

James M Robins, Miguel Angel Hernan, and Babette Brumback. Marginal structural models and causal inference in epidemiology, 2000.

J.M. Robins. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. In L. Sechrest, H. Freeman, and A. Mulley, editors, *Health Service Research Methodology: A Focus on AIDS*, pages 113–159. NCHSR, U.S. Public Health Service, Washington, D.C., 1989.

Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

PR Rosenbaum. Observational studies. *Springer. First citation in articleRosenbaum, PR, & Rubin, DB (1983). The central role of the propensity score in observational studies for causal effects. Biometrika*, 70(1):41–55, 2002.

Evan Rosenman, Guillaume Basse, Art Owen, and Michael Baiocchi. Combining observational and experimental datasets using shrinkage estimators, 2020.

Kangrui Ruan and Xuan Di. Learning human driving behaviors with sequential causal imitation learning. In *Proceedings of the 36nd AAAI Conference on Artificial Intelligence*, 2022.

Kangrui Ruan, Junzhe Zhang, Xuan Di, and Elias Bareinboim. Causal imitation learning via inverse reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023.

D.B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66:688–701, 1974.

Stuart J Russell. *Artificial intelligence a modern approach*. Pearson Education, Inc., 2010.

Stuart J Russell and Peter Norvig. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,, 2016.

Ross D Shachter. Evaluating influence diagrams. *Operations research*, 34(6):871–882, 1986.

Hanif D Sherali and Warren P Adams. *A reformulation-linearization technique for solving discrete and continuous nonconvex problems*, volume 31. Springer Science & Business Media, 2013.

I. Shpitser and J. Pearl. Identification of joint interventional distributions in recursive semi-Markovian causal models. In *Proceedings of the Twenty-First National Conference on Artificial Intelligence*, pages 1219–1226. AAAI Press, Menlo Park, CA, 2006a.

Ilya Shpitser and Judea Pearl. Identification of joint interventional distributions in recursive semi-markovian causal models. In *Proceedings of the National Conference on Artificial Intelligence*, volume 21, page 1219. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2006b.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016.

Satinder P Singh, Tommi Jaakkola, and Michael I Jordan. Learning without state-estimation in partially observable markovian decision processes. In *Machine Learning Proceedings 1994*, pages 284–292. Elsevier, 1994.

Edward Jay Sondik. *The optimal control of partially observable Markov processes*. Stanford University, 1971.

P. Spirtes, C.N. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, Cambridge, MA, 2nd edition, 2001.

Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*. MIT press, 2000.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

R.H. Strotz and H.O.A. Wold. Recursive versus nonrecursive systems: An attempt at synthesis. *Econometrica*, 28:417–427, 1960.

Jos F Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization methods and software*, 11(1-4):625–653, 1999.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015.

Gokul Swamy, Sanjiban Choudhury, Drew Bagnell, and Steven Wu. Causal imitation learning under temporally correlated noise. In *International Conference on Machine Learning*, pages 20877–20890. PMLR, 2022.

Umar Syed and Robert E Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in neural information processing systems*, pages 1449–1456, 2008.

Csaba Szepesvári. *Algorithms for Reinforcement Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2010.

Matthew E Taylor and Peter Stone. Transfer learning for reinforcement learning domains: A survey. *Journal of Machine Learning Research*, 10(Jul):1633–1685, 2009.

Guy Tennenholtz, Uri Shalit, Shie Mannor, and Yonathan Efroni. Bandits with partially observable confounded data. In *Uncertainty in Artificial Intelligence*, pages 430–439. PMLR, 2021.

The White House, Office of the Press Secretary. Fact sheet: Invest in us: The white house summit on early childhood education. Press Release, December 2014.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294, 1933.

J. Tian. *Studies in Causal Reasoning and Learning*. PhD thesis, Computer Science Department, University of California, Los Angeles, CA, November 2002.

J. Tian and J. Pearl. On the testable implications of causal models with hidden variables. In A. Darwiche and N. Friedman, editors, *Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence*, pages 519–527. Morgan Kaufmann, San Francisco, CA, 2002a.

Jin Tian. Identifying conditional causal effects. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 561–568, 2004.

Jin Tian. Identifying dynamic sequential plans. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, pages 554–561, Arlington, Virginia, United States, 2008. AUAI Press. ISBN 0-9749039-4-9.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Proceedings of the Eighteenth National Conference on Artificial Intelligence (AAAI 2002)*, pages 567–573, Menlo Park, CA, 2002b. AAAI Press/The MIT Press.

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

Hoang Tuy, Tuy Hoang, Tuy Hoang, Viêt-nam Mathématicien, Tuy Hoang, and Vietnam Mathematician. *Convex analysis and global optimization*. Springer, 1998.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.

U.S. Department of Health and Human Services. The health consequences of smoking — 50 years of progress: A report of the surgeon general. Technical report, U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, Atlanta, GA, 2014. URL `https://www.surgeongeneral.gov/library/reports/50-years-of-progress/full-report.pdf`.

Benito van der Zander and Maciej Liśkiewicz. Finding minimal d-separators in linear time and applications. In *Uncertainty in Artificial Intelligence*, pages 637–647. PMLR, 2020.

Benito van der Zander, Maciej Liśkiewicz, and Johannes Textor. Constructing separators and adjustment sets in ancestral graphs. In *Proceedings of the UAI 2014 Conference on Causal Inference: Learning and Prediction-Volume 1274*, pages 11–24, 2014.

Hayato Waki, Sunyoung Kim, Masakazu Kojima, Masakazu Muramatsu, and Hiroshi Sugimoto. Algorithm 883: Sparsepop—a sparse semidefinite programming relaxation of polynomial optimization problems. *ACM Transactions on Mathematical Software (TOMS)*, 35(2):15, 2008.

Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. Provably efficient causal reinforcement learning with confounded observational data. *Advances in Neural Information Processing Systems*, 34: 21164–21175, 2021.

Lu Wang, Andrea Rotnitzky, Xihong Lin, Randall E Millikan, and Peter F Thall. Evaluation of viable dynamic treatment regimes in a sequentially randomized trial of advanced prostate cancer. *Journal of the American Statistical Association*, 107(498):493–508, 2012.

Yuhao Wang, Liam Solus, Karren Yang, and Caroline Uhler. Permutation-based causal inference algorithms with interventions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/275d7fb2fd45098ad5c3ece2ed4a2824-Paper.pdf`.

Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. Inequalities for the l1 deviation of the empirical distribution. *Hewlett-Packard Labs, Tech. Rep*, 2003.

B Widrow. Pattern-recognizing control systems. *Compurter and Information Sciences*, 1964.

Xingrui Yu, Yueming Lyu, and Ivor Tsang. Intrinsic reward driven imitation learning via generative model. In *International Conference on Machine Learning*, pages 10925–10935. PMLR, 2020.

J. Zhang and E. Bareinboim. Can humans be out of the loop? Technical Report R-64, Causal Artificial Intelligence Lab, Columbia University, 2020. Also, to appear: Proc. of the 1st Conference on Causal Learning and Reasoning (CLeaR), forthcoming, 2022.

Jiji Zhang. Causal reasoning with ancestral graphs. *Journal of Machine Learning Research*, 9: 1437–1474, 2008.

Junzhe Zhang and Elias Bareinboim. Transfer learning in multi-armed bandits: a causal approach. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 1340–1346. AAAI Press, 2017.

Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In *Advances in Neural Information Processing Systems*, pages 13401–13411, 2019.

Junzhe Zhang and Elias Bareinboim. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *International Conference on Machine Learning*, pages 11012–11022. PMLR, 2020.

Junzhe Zhang and Elias Bareinboim. Bounding causal effects on continuous outcomes. In *Proceedings of the 35nd AAAI Conference on Artificial Intelligence*, 2021.

Junzhe Zhang, Daniel Kumor, and Elias Bareinboim. Causal imitation learning with unobserved confounders. *Advances in Neural Information Processing Systems*, 33, 2020.

Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pages 1433–1438. Chicago, IL, USA, 2008.

Appendix A. Comparison with Partially Observed MDPs

In this section, we will extend the Causal Hierachy Theorem (CHT, (Bareinboim et al., 2020, Thm. 1)) to a general family of stochastic processes where the Markov property does not hold for states and actions across time steps. Specifically, we focus on the partially observed Markov decision processes (POMDP, Sondik (1971)) that model the dependence between observations over a long sequence of time steps through latent states and dynamics. Formally,

Definition 33 (Standard POMDP (Sondik, 1971)) *A partially observable Markov decision process is tuple $\langle \mathscr{D}(S), \mathscr{D}(X), \mathscr{D}(O), \mathcal{T}, \mathcal{R}, \mathcal{O} \rangle$ where*

1. *$\mathscr{D}(S), \mathscr{D}(X), \mathcal{T}$ and $\mathcal{R}$ describe a Markov decision process;*
2. *$\mathscr{D}(O)$ is a finite set of observations the agent can perceive of its world, called the observation space;*
3. *$\mathcal{O}(x, s, o)$ is the observation function, which gives, for each action $X_i = x$ and resulting state $S_{i+1} = s$, a probability distribution over possible observations $O_{i+1} = o$.*

A policy π in a standard POMDP is a sequence of stochastic decision rules $\{\pi_1, \pi_2, \dots\}$; each decision rule π_i is a function mapping from the observations and actions history $\bar{\boldsymbol{O}}_{1:i}, \bar{\boldsymbol{X}}_{1:i-1}$ to a probability distribution over the action space $\mathcal{D}(X)$. Given a policy π and a distribution over the initial state and observation $P(S_1, O_1)$, every standard PMDP model defines a joint distribution over observations $\bar{\boldsymbol{O}}_{1:H}$, actions $\bar{\boldsymbol{X}}_{1:H}$, and rewards $\bar{\boldsymbol{Y}}_{1:H}$ up to decision horizon H, i.e.,

$$\begin{aligned} &P_\pi(\bar{\boldsymbol{o}}_{1:H}, \bar{\boldsymbol{x}}_{1:H}, \bar{\boldsymbol{y}}_{1:H}) \\ &= \sum_{\bar{\boldsymbol{s}}_{1:H}} P(s_1, o_1) \prod_{i=1}^{H} \pi(x_i \mid s_i) \mathcal{T}(s_i, x_i, s_{i+1}) \mathcal{O}(x_i, s_{i+1}, o_{i+1}) \mathbb{1}\{\mathcal{R}(s_i, x_i) = y_i\} \end{aligned} \tag{508}$$

Due to the presence of latent states, the Markov property no longer holds with regard to the perceived observations. This argument is corroborated with the network structure in the causal diagram $\mathcal{G}_{\text{POMDP}}$ of Fig. 10e. For every stage $i = 1, 2, \dots$, conditioning on the observation O_i and action X_i fails to block all paths from observed history O_j, X_j, Y_j for $j < i$ to any future observation O_k, action X_k, and reward Y_k for $k > i$ (Def. 7). Specifically, such long-sequence dependency is generated from the latent states S_i, e.g., the open causal path $O_1 \leftarrow S_1 \rightarrow S_2 \rightarrow O_2$, which could be represented using a standard POMDP.

Example 69 (POMDP, Observational) *Consider the following SCM environment $\mathcal{M}^*$ adapted from Eq. 5, unrolling over stages $i = 1, 2, \dots$*

$$\mathcal{M}^* = \langle \boldsymbol{U} = \{U_{i,1}, U_{i,2}, U_{i,3}\}, \boldsymbol{V} = \{X_i, Y_i, S_i, O_i\}, \mathscr{F} = \{\mathscr{F}_i^*\}, P^*(\boldsymbol{U}) \rangle_{i=1,2,\dots}. \tag{509}$$

The above SCM is identical to the model $\mathcal{M}^*$ defined in Eq. 5, except that the underlying state S_i is now latent to the learner; and the endogenous variables include an observation O_i fixed at a constant $O_i \leftarrow 0$. This means that the learning agent, interacting with the environment, accesses samples drawn from marginal observational $P(\bar{\boldsymbol{o}}_{1:H}, \bar{\boldsymbol{x}}_{1:H}, \bar{\boldsymbol{y}}_{1:H})$ or interventional distribution $P_\pi(\bar{\boldsymbol{o}}_{1:H}, \bar{\boldsymbol{x}}_{1:H}, \bar{\boldsymbol{y}}_{1:H})$, depending on the regimes of interactions.

We compute the observational distributions $P\left(S_{i+1} \mid \bar{\boldsymbol{O}}_{1:i}, \bar{\boldsymbol{X}}_{1:i}\right)$ and $\mathbb{E}\left[Y_i \mid \bar{\boldsymbol{O}}_{1:i}, \bar{\boldsymbol{X}}_{1:i}\right]$ and summarize them using a standard POMDP $\langle \mathscr{D}(S), \mathscr{D}(X), \mathscr{D}(O), \mathcal{T}_{obs}, \mathcal{R}_{obs}, \mathcal{O} \rangle$ where

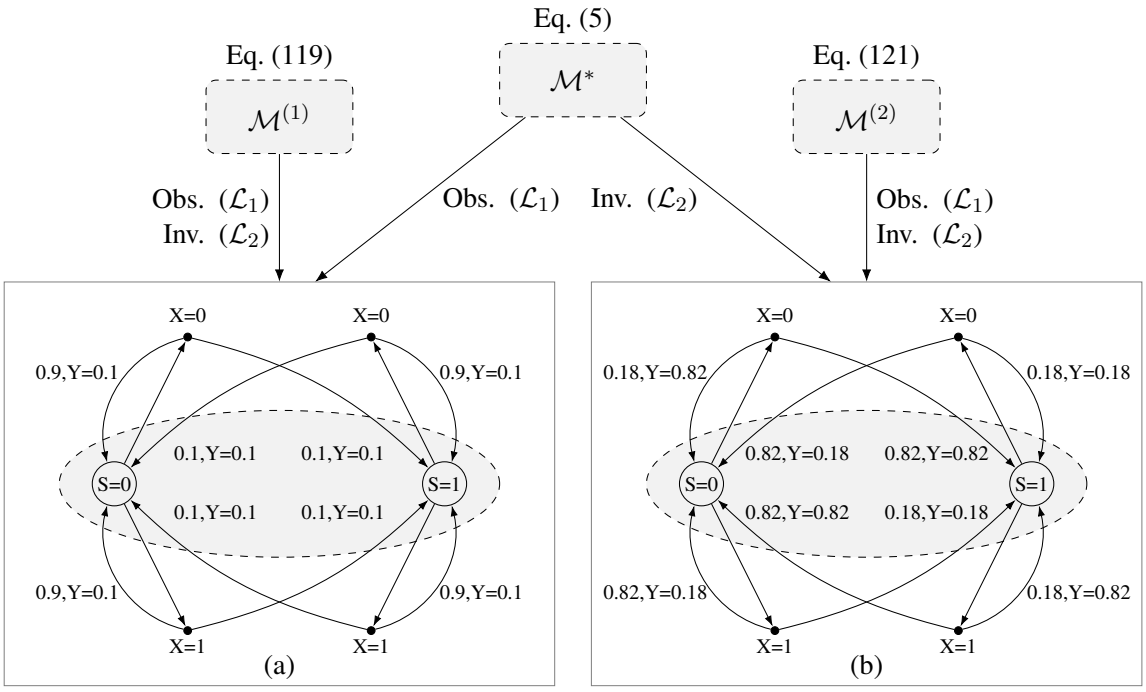


Figure 54: Causal Hierarchy Theorem (CHT) in POMDP environments.

$\mathscr{D}(S), \mathscr{D}(X), \mathcal{T}_{obs}, \mathcal{R}_{obs}$ form a standard MDP described in Example 21; the observation function $\mathcal{O}(x, s, o) = 1$ for observation $o = 0$ given any action $X_i = x$ and subsequent state $S_{i+1} = s$. Fig. 54 (a) shows a finite automaton that describes its detailed system dynamics. The shaded ellipse around the states $S = 0$ and $S = 1$ indicates that both states yield the same observation. ■

Following a similar argument, we could show that any interventional distribution evaluated in a SCM $\mathcal{M}^*$ graphically described in Fig. 10e violates the Markov property, leading to an alternative standard POMDP representation.

Example 70 (POMDP, Interventional) *Consider again the SCM $\mathcal{M}^*$ described in Example 69. Its interventional distributions $P_{\bar{\boldsymbol{X}}_{1:i}}\left(S_{i+1} \mid \bar{\boldsymbol{O}}_{1:i}\right)$ and $\mathbb{E}_{\bar{\boldsymbol{X}}_{1:i}}\left[Y \mid \bar{\boldsymbol{O}}_{1:i}\right]$ a standard POMDP $\langle \mathscr{D}(S), \mathscr{D}(X), \mathscr{D}(O), \mathcal{T}_{exp}, \mathcal{R}_{exp}, \mathcal{O} \rangle$ where $\mathscr{D}(S), \mathscr{D}(X), \mathcal{T}_{exp}, \mathcal{R}_{exp}$ are described in the standard MDP of Example 22; the observation function $\mathcal{O}(x, s, o = 0) = 1$ given any action $X_i = x$ and state $S_{i+1} = s$. The system dynamics of this POMDP are described in the finite automaton of Fig. 54 (b). The ellipse around the states indicates that both states yield the same observation.* ■

In both examples above, the observational and interventional distributions evaluated in the SCM $\mathcal{M}^*$ could be represented using standard POMDPs. However, the detailed system dynamics in these standard POMDPs differ, as illustrated in the finite automata shown in Fig. 54. One may wonder if it is possible to recover interventional quantities $\mathcal{T}_{\text{exp}}$ and $\mathcal{R}_{\text{exp}}$ from the observational data in POMDP environments. Our next result shows this is not the case.

Proposition 8 *For any SCM $\mathcal{M}^*$ compatible with the causal diagram $\mathcal{G}_{\text{POMDP}}$ of Fig. 10e, there is an SCM $\mathcal{M}^{(1)}$ compatible with $\mathcal{G}_{\text{POMDP}}$ such that for every stage $i = 1, 2, \ldots,$*

$$P^{(1)}\left(o_{i+1} \mid \bar{\boldsymbol{o}}_{1:i}, \bar{\boldsymbol{x}}_{1:i}\right) = P^*\left(o_{i+1} \mid \bar{\boldsymbol{o}}_{1:i}, \bar{\boldsymbol{x}}_{1:i}\right), \quad \mathbb{E}^{(1)}\left[Y_i \mid \bar{\boldsymbol{o}}_{1:i}, \bar{\boldsymbol{x}}_{1:i}\right] = \mathbb{E}^*\left[Y_i \mid \bar{\boldsymbol{o}}_{1:i}, \bar{\boldsymbol{x}}_{1:i}\right] \quad (510)$$

while

$$P^{(1)}_{\bar{\boldsymbol{x}}_{1:i}}\left(o_{i+1} \mid \bar{\boldsymbol{o}}_{1:i}\right) \neq P^{*}_{\bar{\boldsymbol{x}}_{1:i}}\left(o_{i+1} \mid \bar{\boldsymbol{o}}_{1:i}\right), \qquad \mathbb{E}^{(1)}_{\bar{\boldsymbol{x}}_{1:i}}\left[Y_i \mid \bar{\boldsymbol{o}}_{1:i}\right] \neq \mathbb{E}^{*}_{\bar{\boldsymbol{x}}_{1:i}}\left[Y_i \mid \bar{\boldsymbol{o}}_{1:i}\right] \tag{511}$$

■

The following example constructs an alternative SCM $\mathcal{M}^{(1)}$ that generates the observational distribution as the underlying environment $\mathcal{M}^{*}$, but differs significantly in interventional distributions.

Example 71 (POMDP, Observational $\not\Rightarrow$ Interventional) *We will construct an alternative SCM $\mathcal{M}^{(1)}$ where its system dynamics collapse to the associational layer ($\mathcal{L}_1$) with respect to the SCM $\mathcal{M}^{*}$ described in Example 69. More specifically,*

$$\mathcal{M}^{(1)} = \left\langle \boldsymbol{U} = \{U_{i,1}, U_{i,2}, U_{i,3}\}, \boldsymbol{V} = \{X_i, Y_i, S_i, O_i\}, \mathscr{F} = \left\{\mathscr{F}_i^{(1)}\right\}, P^{(1)}(\boldsymbol{U})\right\rangle_{i=1,2,\ldots}. \tag{512}$$

Similarly to the construction of $\mathcal{M}^{*}$, the above SCM is identical to the model $\mathcal{M}^{(1)}$ defined in Example 23 except that for every stage $i = 1, 2, \ldots$, the learner does not observe the state S_i, but only receives new endogenous variable $O_i \leftarrow 0$. We compute the observational distributions $P\left(S_{i+1} \mid \bar{\boldsymbol{O}}_{1:i}, \bar{\boldsymbol{X}}_{1:i}\right)$ and $\mathbb{E}\left[Y_i \mid \bar{\boldsymbol{O}}_{1:i}, \bar{\boldsymbol{X}}_{1:i}\right]$ and the interventional distributions $P_{\bar{\boldsymbol{X}}_{1:i}}\left(S_{i+1} \mid \bar{\boldsymbol{O}}_{1:i}\right)$ and $\mathbb{E}_{\bar{\boldsymbol{X}}_{1:i}}\left[Y \mid \bar{\boldsymbol{O}}_{1:i}\right]$ evaluated in $\mathcal{M}^{(1)}$, following the discussion in Example 23. The analysis suggests that the observational and interventional distributions collapse in the model $\mathcal{M}^{(1)}$, which could be described using the same finite-state automaton shown in Fig. 54(a).

Compared with system dynamics in $\mathcal{M}^{*}$, the model $\mathcal{M}^{(1)}$ coincides with $\mathcal{M}^{*}$ in the observational distributions ($\mathcal{L}_1$), but deviates significantly in the interventional distributions ($\mathcal{L}_2$). More specifically, given any observations and actions' history $\bar{\boldsymbol{o}}_{1:i}, \bar{\boldsymbol{x}}_{1:i}$, the induced reward function in $\mathcal{M}^{(1)}$ is given by,

$$\begin{aligned} \mathbb{E}^{(1)}_{\bar{\boldsymbol{x}}_{1:i}}\left[Y_i \mid \bar{\boldsymbol{o}}_{1:i}\right] &= \mathbb{E}^{(1)}\left[Y_i \mid \bar{\boldsymbol{o}}_{1:i}, \bar{\boldsymbol{x}}_{1:i}\right] && (513)\\ &= 0.1 && (514)\end{aligned}$$

On the other hand, suppose the state occupancy rate $P(s_i) = 0.5$ for any state $s_i = 0, 1$ in the model $\mathcal{M}^{*}$. Evaluating the transition distribution in $\mathcal{M}^{*}$ gives, for any action $x_i = 0, 1$ and any state $s_{i+1} = 0, 1$,

$$P_{x_i}\left(s_{i+1}\right) = 0.5. \tag{515}$$

For instance, let $x_i = 0$ and $s_{i+1} = 1$. Expanding on the current state S_i gives,

$$\begin{aligned} &P_{X_i \leftarrow 0}\left(S_{i+1} = 1\right) && (516)\\ &= P_{X_i \leftarrow 0}\left(S_{i+1} = 1 \mid S_i = 0\right) P(S_i = 0) + P_{X_i \leftarrow 0}\left(S_{i+1} = 1 \mid S_i = 1\right) P(S_i = 1) && (517)\\ &= 0.82 \times 0.5 + 0.18 \times 0.5 && (518)\\ &= 0.5 && (519)\end{aligned}$$

The above equations imply the following intermediate reward evaluated in the model $\mathcal{M}^{*}$

$$\begin{aligned} \mathbb{E}^{*}_{\bar{\boldsymbol{x}}_{1:i}}\left[Y_i \mid \bar{\boldsymbol{o}}_{1:i}\right] &= \sum_{s_i} \mathbb{E}^{*}_{x_i}\left[Y_i \mid s_i\right] P^{*}_{\bar{\boldsymbol{x}}_{1:i-1}}(s_i \mid \bar{\boldsymbol{o}}_{1:i}) && (520)\\ &= 0.5, && (521)\end{aligned}$$

which deviates from the corresponding query in $\mathcal{M}^{(1)}$. For instance, let $x_i = 0$. Expanding on the current state S_i gives,

$$\mathbb{E}^*_{\bar{\boldsymbol{x}}_{1:i}}[Y_i \mid \bar{\boldsymbol{o}}_{1:i}] \tag{522}$$
$$= \mathbb{E}^*_{X_i \leftarrow 0}[Y_i \mid S_i = 0]\, P^*_{\bar{\boldsymbol{x}}_{1:i-1}}(S_i = 0 \mid \bar{\boldsymbol{o}}_{1:i}) + \mathbb{E}^*_{X_i \leftarrow 0}[Y_i \mid S_i = 1]\, P^*_{\bar{\boldsymbol{x}}_{1:i-1}}(S_i = 1 \mid \bar{\boldsymbol{o}}_{1:i}) \tag{523}$$
$$= 0.82 P^*_{\bar{\boldsymbol{x}}_{1:i-1}}(S_i = 0 \mid \bar{\boldsymbol{o}}_{1:i}) + 0.18 P^*_{\bar{\boldsymbol{x}}_{1:i-1}}(S_i = 1 \mid \bar{\boldsymbol{o}}_{1:i}) \tag{524}$$
$$= 0.5 \tag{525}$$

The last step holds since the occupancy rate $P^*_{\bar{\boldsymbol{x}}_{1:i-1}}(s_i \mid \bar{\boldsymbol{o}}_{1:i}) = 0.5$ for any state $s_i = 0, 1$. This example corroborates Prop. 8 and illustrates that observational queries are generally underdetermined by randomized experiments in POMDP environments. ■

The above example shows that interventional distributions in an unknown POMDP environment are generally not fully determined from the observational distribution. Conversely, we also show that it is generally infeasible to recover observational quantities from randomized experiments in non-Markov processes.

Proposition 9 *For any SCM $\mathcal{M}^*$ compatible with the causal diagram $\mathcal{G}_{\text{POMDP}}$ of Fig. 10e, there is an SCM $\mathcal{M}^{(2)}$ compatible with $\mathcal{G}_{\text{POMDP}}$ such that for every stage $i = 1, 2, \ldots,$*

$$P^{(2)}_{\bar{\boldsymbol{x}}_{1:i}}(o_{i+1} \mid \bar{\boldsymbol{o}}_{1:i}) = P^*_{\bar{\boldsymbol{x}}_{1:i}}(o_{i+1} \mid \bar{\boldsymbol{o}}_{1:i}), \qquad \mathbb{E}^{(2)}_{\bar{\boldsymbol{x}}_{1:i}}[Y_i \mid \bar{\boldsymbol{o}}_{1:i}] = \mathbb{E}^*_{\bar{\boldsymbol{x}}_{1:i}}[Y_i \mid \bar{\boldsymbol{o}}_{1:i}] \tag{526}$$

while

$$P^{(1)}(o_{i+1} \mid \bar{\boldsymbol{o}}_{1:i}, \bar{\boldsymbol{x}}_{1:i}) \neq P^*(o_{i+1} \mid \bar{\boldsymbol{o}}_{1:i}, \bar{\boldsymbol{x}}_{1:i}), \quad \mathbb{E}^{(1)}[Y_i \mid \bar{\boldsymbol{o}}_{1:i}, \bar{\boldsymbol{x}}_{1:i}] \neq \mathbb{E}^*[Y_i \mid \bar{\boldsymbol{o}}_{1:i}, \bar{\boldsymbol{x}}_{1:i}] \tag{527}$$

■

The following example corroborates the proposition mentioned above by constructing an alternative SCM $\mathcal{M}^{(2)}$ compatible with the causal diagram of Fig. 10e that induces the same interventional distribution in the underlying environment but generates different observations.

Example 72 (POMDP, Interventional $\not\Rightarrow$ Observational) *Consider the following SCM environment rolling over stages $i = 1, 2, \ldots,$*

$$\mathcal{M}^{(2)} = \left\langle \boldsymbol{U} = \{U_{i,1}, U_{i,2}, U_{i,3}\}, \boldsymbol{V} = \{X_i, Y_i, S_i, O_i\}, \mathscr{F} = \left\{\mathscr{F}_i^{(2)}\right\}, P^{(2)}(\boldsymbol{U})\right\rangle_{i=1,2,\ldots}. \tag{528}$$

Similar to the previous example, the above SCM is identical to the model $\mathcal{M}^{(2)}$ defined in Example 24 except that for every stage $i = 1, 2, \ldots,$ the learner does not observe the state S_i, but only receives new endogenous variable $O_i \leftarrow 0$. We compute the observational distributions $P\left(S_{i+1} \mid \bar{\boldsymbol{O}}_{1:i}, \bar{\boldsymbol{X}}_{1:i}\right)$ and $\mathbb{E}\left[Y_i \mid \bar{\boldsymbol{O}}_{1:i}, \bar{\boldsymbol{X}}_{1:i}\right]$ and the interventional distributions $P_{\bar{\boldsymbol{X}}_{1:i}}\left(S_{i+1} \mid \bar{\boldsymbol{O}}_{1:i}\right)$ and $\mathbb{E}_{\bar{\boldsymbol{X}}_{1:i}}\left[Y \mid \bar{\boldsymbol{O}}_{1:i}\right]$ evaluated in $\mathcal{M}^{(2)}$, following the discussion in Example 24. The analysis suggests that the observational and interventional distributions coincide in the model $\mathcal{M}^{(2)}$, which could be described using the same finite automaton of Fig. 54(b).

Comparing with system dynamics in the SCM $\mathcal{M}^*$ described in Example 71, we find that model $\mathcal{M}^{(2)}$ coincides with $\mathcal{M}^*$ in the interventional distributions ($\mathcal{L}_2$), but disagree in the observational

distributions ($\mathcal{L}_1$). More specifically, given any observations and actions' history $\bar{\boldsymbol{o}}_{1:i}, \bar{\boldsymbol{x}}_{1:i}$, the observed intermediate reward in $\mathcal{M}^{(2)}$ is given by,

$$\mathbb{E}^{(2)}\left[Y_i \mid \bar{\boldsymbol{o}}_{1:i}, \bar{\boldsymbol{x}}_{1:i}\right] = \mathbb{E}^{(2)}_{\bar{\boldsymbol{x}}_{1:i}}\left[Y_i \mid \bar{\boldsymbol{o}}_{1:i}\right] \tag{529}$$

$$= 0.5 \tag{530}$$

On the other hand, evaluating the observed intermediate reward given the history $\bar{\boldsymbol{o}}_{1:i}, \bar{\boldsymbol{x}}_{1:i}$ in the model $\mathcal{M}^*$ gives

$$\mathbb{E}^*\left[Y_i \mid \bar{\boldsymbol{o}}_{1:i}, \bar{\boldsymbol{x}}_{1:i}\right] = 0.1, \tag{531}$$

which differs from the corresponding query in $\mathcal{M}^{(2)}$. This complements previous examples and illustrates that interventional queries are generally non-identifiable from the observational data in POMDP environments. ■

We organize the examples and results discussed in this section and summarize them in Fig. 54. The agent interacts with the ground-truth SCM $\mathcal{M}^*$ (Example 71) in the middle, through passive observation or active intervention, and generates the observational and interventional distributions. The system dynamics of these distributions do not satisfy the Markov property and can be represented using the latent states in standard POMDPs. The finite automata in Fig. 54 (a, b) describes these latent dynamics, respectively.

Assuming only observational data ($\mathcal{L}_1$) is available, one can construct an alternative SCM $\mathcal{M}^{(1)}$ (left side) that generates the same the observational data but have different interventional distributions (i.e., $\mathcal{L}_1^* = \mathcal{L}_1^{(1)}$, $\mathcal{L}_2^* \neq \mathcal{L}_2^{(1)}$). This implies that, in practice, natural trajectories of other behavioral agents collected from passive observations are generally insufficient to make claims about the learning agent's actions and performance. Conversely, whenever the interventional data ($\mathcal{L}_2$) is available, one can construct an alternative SCM $\mathcal{M}^{(2)}$ (right side) that generates the same interventional distribution but has a different observational one (i.e., $\mathcal{L}_1^* \neq \mathcal{L}_1^{(2)}$, $\mathcal{L}_2^* = \mathcal{L}_2^{(2)}$). This might seem counterintuitive, as interventions are generally thought to be more informative than simply observing a system as it evolves over time. However, in practice, this approach does not enable the learning agent to predict how other agents will behave within the same environment.