Climbing the Ladder: A Survey of Counterfactual Methods in Decision Making Processes

St John M.M. Grimbly

GRMSTJ001
Supervisor: Dr Jonathan Shock
University of Cape Town

November 1, 2020

Abstract

The field of reinforcement learning (RL) has seen a surge of interest due to its recent successes in various sequential decision making domains. RL is concerned with maximisation of cumulative reward over long time horizons. On the other hand, causal inference provides a set of tools and techniques to combine structural information about the data generating process, and data itself, to reason and infer up to a counterfactual nature - *what would have happened had something been different?* By adding causal structural information to sample efficient RL techniques we can boost learning performance in many domains. This combination of theory from different fields has led to successes in various domains. Elias Bareinboim encapsulates these as *The Six Tasks* of Causal Reinforcement Learning (CRL). This paper develops, surveys, discusses and brings together much of the recent work in this emerging field, within the context and framework of developing and working towards generally intelligent agents.

Contents

1 Introduction

"*All reasonings concerning matter of fact seem to be founded on the relation of cause and effect. By means of that relation alone we can beyond the evidence of our memory and senses.*" - David Hume [1].

What's the first thing a statistician will say when you dare say the word cause? If you've ever taken a statistics class, I have little doubt it was the classic anachronism, *correlation does not imply causation.* This trope seems to imply that we can never state that A causes B by analysis of data alone. R.A. Fisher, one of the fathers of modern statistics, was stringently opposed to causal conclusions without overwhelming evidence. See, for example, the debate on whether smoking causes cancer (e.g. [2], [3], [4]). But is statistics really powerless when faced with the battle of determining causation? Recently there has been much interest in the techniques developed to infer causation from data, especially with the proliferation of statistical learning techniques in the fields of deep learning, for example. Graphical techniques employing DAGs have been especially popular in recent years, often locking horns with those in favour of the Neyman–Rubin causal model [5] and potential outcomes framework (e.g. [6]). This *causal revolution* is exactly what Judea Pearl's popular science book, *The Book of Why* [7], among other popular works, addresses.

With the rise in popularity of unsupervised methods in both machine learning and reinforcement learning paradigms, there is little doubt that inferring causal structure will play a crucial role in getting artificial agents to make informed decisions, especially in an uncertain world. Reinforcement learning agents are concerned with maximising cumulative reward over a long time horizon by following a sequence of optimal actions. Ensuring such an agent maintains a causal model for the world it operates in will undoubtedly make for interpretable models in a field otherwise filled with 'black-boxes'. One should be careful not to confuse *world models* in RL with a causal model. A causal model explicitly models the nature of the relationships of the underlying data generating process, whereas a RL world model attempts to simulate the predictive outputs due to an agent intervention. Causal and graphical models also extend the applicability of current decision making methods, most of which are only applicable under a narrow set of assumptions. Consider that reinforcement learning under an MDP formulation explicitly requires Markov processes. As we will discuss, this fails to account for some fundamental decision processes in the real world - including dynamic treatment regimes in the field of personalised medicine. One can easily imagine how a generally intelligent agent would require long-term, non-Markov planning ability.

Before continuing to develop the notions of causality, it should be made clear that we are working within the *ladder of causation* proposed by Judea Pearl [8]. In this framework there are three rungs on the causal hierarchy, each adding additional information not available to models belonging to a lower rung. These are (1) observational information, (2) interventional information, and (3) counterfactual information, with each building upon and subsuming the last. Reinforcement learning naturally falls on the interventional rung since agents learn about optimal actions by observing outcomes due to their interventions in the system. They cannot, however, use interventional data to answer counterfactual - "what if?" - style questions without additional information. This is critical to acknowledge for much of the theory that follows. The astute reader will point out that counterfactual quantities are inherently non-scientific because they cannot be proved to be true - what has happened, has happened. We do not claim otherwise. Rather, counterfactual quantities are useful to for decision making - as is clear with any thought about how we reason in our daily lives.

This paper will introduce the mathematical notion of causality from the perspective of statistics, and place it in the context of machine learning and artificial intelligence. Specifically, focus will be placed on developing and discussing the theory of causal reinforcement learning (CRL) so that an interested reader is prepared for dealing with state-of-the-art research and results. The six tasks introduced by Bareinboim [9] will be discussed through the surveying of relevant and recent literature. Finally the state of causal reinforcement will be discussed in the context of the current machine learning landscape and the quest for artificial general intelligence (AGI).

2 Notation

In this section we briefly introduce the notation used throughout this paper. For the most part we stick to conventional statistical notation found in the literature. Where this does not hold, notational complexities will be explained. Random variables are presented with a capital letter, X. A realisation of this variable is displayed in lowercase x. Boldface is usually reserved for a set, $\boldsymbol{X}$. A sequence, when otherwise not clear, will be shown with an overline, $\overline{X}$. Calligraphic letters will be used to denote models or special types where it could otherwise be confused with a random variable. For example, a causal model will be written as G, or $\mathcal{G}$ if not clear. Calligraphy will also be used to denote domains when clear, $\mathcal{X}$. Where it is more clear, domains will be explicitly written as $Dom(X)$. Optimal values, variables, or policies are notated with an $*$. $\perp\!\!\!\perp$ is used to denote conditional independence between variables. π will usually be reserved for a policy. Ancestral relationships from graph theory will be used. For example, $Pa(X)$ denotes the parents of X. This will usually differ from $pa(X)$ by including X itself. Where clear, an overline represents the causal model with incoming edges removed, $G_{\overline{X}}$. Similarly an underline for removal of outgoing edges, $G_{\underline{X}}$. $\langle\cdot\rangle$ will denote a tuple where convenient.

3 Preliminaries

As you probably recall from high school, probability and statistics are almost entirely formulated on the idea of drawing random samples from an experiment. One imagines observing realisations of outcomes from some set of possibilities when drawing from an assortment of independent and identically distributed (i.i.d.) events. In reality, this assumption of i.i.d. events fails in many situations. Consider shifting some distribution of events or intervening in the system. This failure of an often fundamental assumption in statistics is one reason for the causal approach we shall develop. In probability theory and statistics, we try to predict an outcome and then associate some probability of an event occurring given some distribution of the underlying space of events. In statistical learning, including machine learning, we are performing the inverse problem. We are trying to find the underlying *description* of the data. In statistics, the likelihood approach for inference is common for inference. Statistical machine learning can be seen as a simple extension of this approach - applying information gathered from collected data to infer patterns or associations which appear due to the data generating process.

The causal inference and causal learning problem is a harder one. Even if we had perfect and complete knowledge of the observational distribution we would still not know the underlying causal structure of the data. Causal modelling is more fundamental than the probabilistic approach since additional information about relationships between variables is contained in such a model. Causal reasoning allows us to analyse the effects of interventions or distribution changes and make predictions in a more general sense than conventional statistical approaches. Further, it allows for counterfactual reasoning - an ability a reinforcement learning agent generally lacks. Inferring causal structure thus becomes the inverse problem. Using observational data and outcomes, as well as intervention data, we would like to infer the underlying relationships between the appropriate variables of interest. These relationships between the different modelling and reasoning methods are visualised in figure 1. Although we cannot infer concrete causal structure, we can at least infer the existence of the underlying causal connections from statistical dependence between them. This paper introduces some advanced theory and techniques for this very subject.

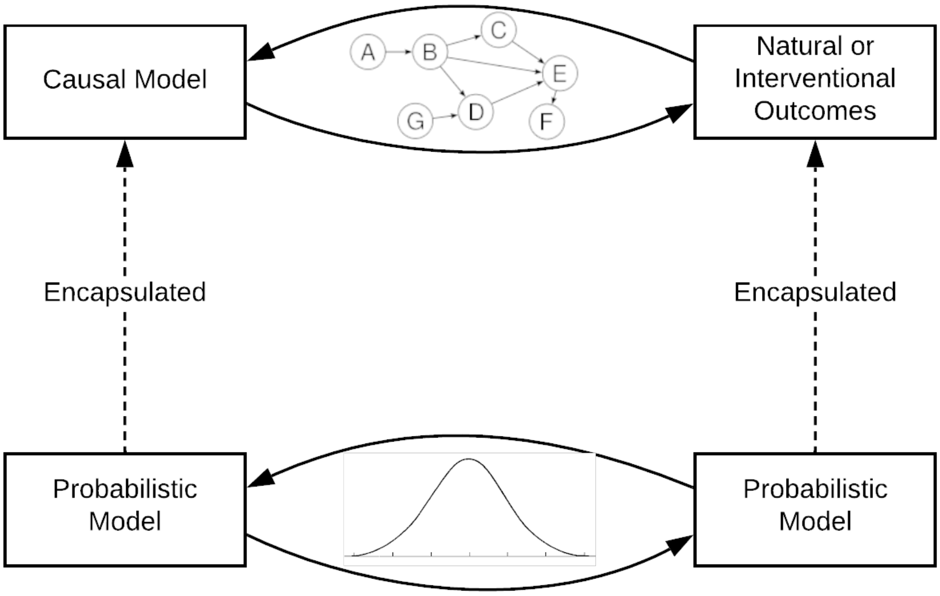


Figure 1: Figure showing the relationship between probabilistic regimes and causal regimes of reasoning and learning. Here we can see that the causal paradigm generalises the conventional statistical approaches to learning and prediction. Reinforcement learning fits in as the intervention step in the causal learning procedure. The idea is to augment RL by adding a model to aid learning and reasoning. Figure augmented from and inspired by [10] - a good resource for introductory causal inference material.

With the preamble done, this section will briefly introduce necessary notions which will either be directly used or helpful in a general context throughout this paper. More detail about basic causal inference theory and reinforcement learning theory is available in the appendices. Perhaps the most fundamental formalism to this formulation of causal theory is the idea of a structural causal model.

Definition 3.1 (Structural Causal Model (SCM) [8]). *A structural causal model M is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, F, P(\mathbf{U}) \rangle$, where*

1. *$\mathbf{U}$ are the exogenous variables, determined entirely by external factors.*

2. *$\mathbf{V}$ are the endogenous variables, determined by other variables in the model.*

3. *F is the set of mappings from $\mathbf{U}$ to $\mathbf{V}$ such that each $f_i \in F$ maps from $U_i \cup Pa_i$ to V_i by $v_i \leftarrow f_i(pa_i, u_i)$. Here $pa_i \in Pa_i$, $U_i \subseteq \mathbf{U}$, and $Pa_i \subseteq \mathbf{V} \setminus V_i$. In other words, it assigns a value dependent on other variables in the model to a specific variable.*

4. *$P(\mathbf{U})$ is a probability density defined over domain of $\mathbf{U}$.*

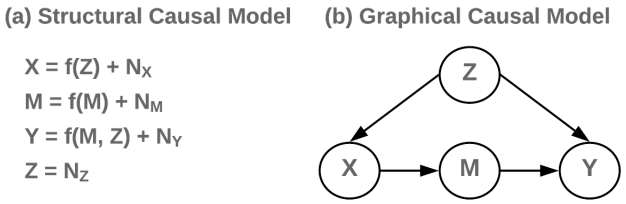


Figure 2: Figure showing an SCM with an associated graphical representation. Here Z is exogenous while the other variables are endogenous. Each variable has associated noise variables, N, which indicate the probabilistic nature of the assignments. Figure self-created.

Another important concept is that of *identifiability.* Notice that an SCM induces a joint distribution over the variables of interest. For example, the SCM $C \rightarrow E$ induces $P_{C,E}$. Naturally, we wonder whether we can *identify*, in general, whether the joint distribution came from the model $C \rightarrow E$ or $E \rightarrow C$. It turns out we cannot since the graphs are not unique in inducing this joint distribution. In other words, structure is not *identifiable* from the joint distribution because the underlying graphs add an additional layer of knowledge to the that given by the joint. Proposition (3.1) formulates this idea by indicating that we can construct an SCM from a joint distribution in any direction - i.e ← or →. This is crucial to keep in mind, especially if we plan on trying to use observational data to infer causal structure.

Proposition 3.1 (Non-uniqueness of graph structures [10]). *For every joint distribution $P_{X,Y}$ of two real-valued variables, there is an SCM*

$$Y = f_Y(X, N_Y), X \perp Y,$$

where f_Y is a measurable function and N_Y is a real-valued noise variable.

Another important property that will be useful in later analysis is that of d-separation. Essentially, this tells us about the conditional independence relations available in the causal model. In some way, this tells us what information (in the form of variables) 'links' other variables by way of a causal path. In later sections we will discover this is a very useful property in causal learning and graph manipulation for some important algorithms.

Definition 3.2 (d-separation [11]). *A set Z of nodes in a causal graph is said to block a path p if either (1) p contains at least one incoming or outgoing edge that traverses a vertex in Z, or (2) p contains at least one collision vertex that is outside Z and has no descendant in Z. If Z blocks all available paths from sets X to Y, the X and Y are said to be d-separated by Z.*

In the context of causal learning, the d-separation property informs us of how variables are dependent on each other. For example, if X and Y are d-separated by Z, we know information cannot travel between X and Y by some *backdoor* if we control for Z. This is a critical notion as it underlies a large portion of this paper dealing with unobserved confounding and latent variables. Another important concept is that of *faithfulness.* This assumption indicates that causal relations are only formed as a result of d-separation.

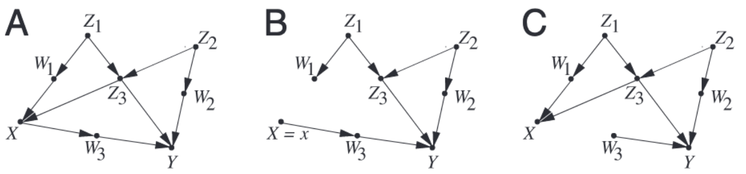


Figure 3: Causal graphs showing examples of theory discussed in this section. (A) shows example of d-separation. (B) shows the resultant graph of an intervention on X, $do(X = x)$. (C) shows the result of trying to block the direct path between X and Y resulting in backdoor paths. Figure extracted from [11].

Judea Pearl presents three rules of *do*-calculus that proves sufficient for a wide variety of manipulations between the 'rungs' on the ladder of causation (see page 234, [7]). The first rule is fairly obvious for the most part. When an observation independent of the variables of interest is made, it does not affect the probability distribution. Formally, $P(Y \mid do(X), Z, W) = P(Y \mid do(X), Z)$ where Z blocks all paths from W to Y in causal model $G_{\overline{X}}$. That is, the model with arrows directed into X removed. The second rule deals with backdoor paths. If Z satisfies the back-door criterion then $P(Y \mid do(X), Z) = P(Y \mid X, Z)$. In other words, after controlling for all necessary confounding factors, observation matches intervention. Finally, if there is no causal relationship between X and Y, then $P(Y \mid do(X)) = P(Y)$. These three simple rules are remarkably effective in a wide range of proofs. They will prove useful in some examples we discuss in later sections.

Ultimately, in this paper we are interested in optimising decision making procedures. The Multi-Armed Bandit (MAB) problem is perhaps the most popular and simplistic setting encountered in literature discussing sequential decision making, and it serves as a key starting point in studies of reinforcement learning (RL). The theory of MAB decision making is also rich and fairly complete for simple scenarios that satisfy strict assumptions. Sutton and Barto [12] - the authors of a seminal introductory text in RL - discuss this problem in detail in the context of optimal control and Bellman optimality. The MAB problem involves maximising the expected reward/payout given that the reward distribution of each bandit arm is initially unknown to the agent. This can be rephrased as minimisation of the *regret* the agent experiences, which is the form often encountered in causal and optimal control literature. The regret of a *policy* or *allocation strategy* A after n plays is defined by

$$\mu^* n - \mu_j \sum_{j=1}^{K} \mathbb{E}[T_j(n)] \quad \text{where } \mu^* = \max_{1 \le i \le K} \mu_i.$$

Here $T_i(n)$ denotes the number of times machine i has been played by policy A, μ represents the expected reward a machine, and $*$ denotes the optimal policy. The regret is thus the expected loss due to suboptimal actions by a policy. Regret is a natural and intuitive quantity to work with and serves as a proxy for efficiency in learning - key to general intelligence for artificial agents. Regret is a key quantity that this work of causal reinforcement learning addresses. Of course, reinforcement learning is usually interested in sequential decision making over a long-term horizon. This is usually formulated in terms of MDPs.

Definition 3.3 (Markov Decision Process (MDP) [12])**.** *A Markov decision process (MDP) is a 5-tuple $\langle S, A, P, R, \gamma \rangle$ of states S, actions A, transition probabilities P, rewards R, and a discount factor γ. Given a state and action, the transition map determines the next state and an associated reward according to the transition probability.*

This MDP model makes several strict, implicit assumptions about the system of interest. For example, that states are fully observed and there is no confounding on any variable of interest. The Markov assumption clearly fails in many areas of interest, including personalised healthcare treatment regimes (dealt with in later sections). This is a key motivator for the need to add causal information to the reinforcement learning theory. We are now ready to begin developing the causal reinforcement learning theory. We start by discussing the context of the six tasks this paper presents.

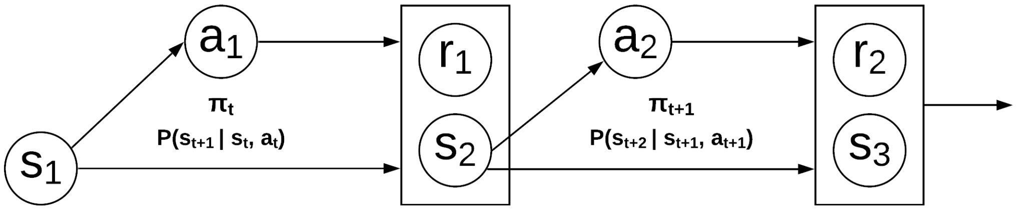

a_1
r_1
a_2
r_2
π_t
$P(s_{t+1} \mid s_t, a_t)$
π_{t+1}
$P(s_{t+2} \mid s_{t+1}, a_{t+1})$
s_1
s_2
s_3

Figure 4: Figure showing the relationships between variables in a classic RL MDP setting. Clearly the process is Markov in nature. States and rewards are solely based on previous states and taken action. The action is guided by the policy, π_t. Figure was created for this paper and loosely inspired by examples used in classic texts. See [13] for a good introduction to RL.

4 Causal Reinforcement Learning: The Six Tasks

Causal inference establishes a set of principles and practices for dealing with data from a structural level. Framing the data generating process in causal language while making assumptions about the underlying generative model explicit allows us to reason about causal relationships in a counterfactual nature. Reinforcement learning is concerned with maximising reward in the face of uncertainty in potentially foreign environments and data domains. These two fields, though seemingly disparate, both deal with data in an interventional and - possibly - counterfactual manner. Elias Bareinboim's CausalAI Lab at Columbia University have sought to tie these fields together by placing them under a single conceptual and theoretical framework [9]. This combination of work has - and is - resulting in powerful results that has not been possible without such an approach. Bareinboim encapsulates the work in this domain as falling under a set of six tasks that these fields can jointly solve and contribute to. He dubs this area of research Causal Reinforcement Learning (CRL). This forms the crux of what this paper investigates, surveys and introduces. It should be noted that Bareinboim et al. are in the process of creating such an introduction to CRL themselves. This paper merely serves to independently survey the field and perhaps provide an alternate context for the framing of this field of work. The interested and motivated reader is encouraged to refer to the source material. Researchers in this area are very active.

We begin by discussing generalised policy learning. In general, this involves systematically combining offline and online modes of observation and interaction (i.e. intervention) with an environment to boost learning performance. We then discuss the problem of identifying when and where to intervene in a causal system. A large portion of this writing is devoted to the third task - counterfactual decision making. This involves exploiting both observational and experimental data to reason about counterfactual quantities and boost learning performance. By leveraging information contained in an agent's intended action we can learn about unseen factors that are influencing and confounding the system. Next we discuss ideas and methods of learning about structural invariances in a causal system to aid in the transportability of data between domains. The fifth task is focused on learning causal structure from observation and interaction with the environment. Finally, we discuss causal imitation learning. The overall approach to this paper is to include proofs where they aid in explanation or provide important insight into a method or approach to a problem. Additionally, where they demonstrate some useful technique or simply deemed interesting, they will be included. Similarly for placement of figures and included algorithms. This paper is written with brevity in mind, but includes discussion where critical and aids to the overall theme of developing theory for working towards general intelligence. We start by discussing some modern methods for generalising policy learning for a combination of online and offline domains over time horizons that are not necessarily Markov - think healthcare.

4.1 Task 1: Generalised Policy Learning

Reinforcement learning typically involves learning and optimising some policy about how to interact in an environment to maximise some reward signal. Typical reinforcement learning agents are trained in isolation, exploiting copious amounts of computing power and energy resources. In a crude manner of speaking, offline policy learning involves learning from a fixed set of collated data. Online policy learning typically involves learning on-the-fly, with the main constraint being time. This approach requires flexibility as data can change over time without any indication of such a change to the agent. In addition, state-of-the-art agents often take a substantial amount of time to train. Transfer learning seeks to solve this inefficiency in the learning process by applying previous knowledge and experience to boost learning performance, similar to how humans can exploit previous knowledge to solve novel tasks. This is discussed in more detail in section 4.6. The field of causal inference similarly deals with this problem of inferring effect from heterogeneous sources of data. A major problem in this process involves learning in the face of unobserved (hidden) confounders. We now discuss some ideas of how causal inference and modelling can be applied to multi-armed bandits (MABs) and Markov decision processes (MDPs) to boost learning performance by combining different modes of learning - observation and interventional.

One such paper that tackles this problem is [14] in which the authors combine transfer learning in (basic) reinforcement learning with causal inference theory. This is done in the context of two multi-armed bandit (MAB) agents given access to a causal model of the environment. In the case where causal effects are not identifiable, the authors provide a method of extracting *causal bounds* from available knowledge contained in the available distributions. We now develop some of the theory presented in this paper.

Contextual bandits are discussed in [15] and are a variation of MABs such that the agent can observe additional information (context) associated with the reward signal. The authors of [14] start by considering an off-policy learning problem such that agent A follows some policy $do(X = \pi(\epsilon, u))$ with context $u \in U$ and noise ϵ, resulting in joint distribution $P(x, y, u)$. Another agent, A', would similarly like to learn about the environment and exploit the experience of A to boost its learning and quickly converging upon the optimal policy. This problem boils down to identifying the causal effect of an intervention on X, given by $\mathbb{E}[Y \mid do(x)]$. That is, the expected outcome given that we experiment by intervening on X. A more challenging scenario appears if we wish to transfer knowledge from this contextual bandit to a standard MAB agent, say B (see figure 5).

If we denote by $G_{\overline{X}\underline{Z}}$ the subgraph obtained by deleting all edges directed into X and all edges directed out of Z, then $(Y \perp\!\!\!\perp X \mid U)_{\underline{G}}$ means that by removing edges directed out of G, given the context U, we obtain that Y is independent of X. This is useful because we can then derive $\mathbb{E}[Y \mid do(x)] = \sum_{u \in D(U)} \mathbb{E}[Y \mid x, u] P(u)$ (applying the second rule of *do*-calculus). In this case the average effect $\mathbb{E}[Y \mid do(x)]$ was identifiable - there were not multiple causal structures inducing the same distribution. The authors note that *do*-calculus provides a complete method for identifying such causal effects but it is not useful for constructing such formulae for non-identifiable queries. For example, if agent B cannot observe the context in which A operates, *do*-calculus cannot identify the average effect $\mathbb{E}[Y \mid do(x)]$ since different causal models can induce the same observational distribution $P(x, y)$ with different expected rewards. This is a very important concept to note since naivete in the transfer process under these conditions can lead to negative impact on the performance of the target. In practice, however, we often don't have access to the underlying SCMs beforehand, and thus cannot distinguish between two such models.

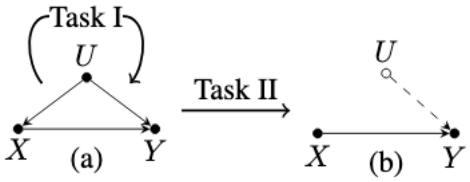


Figure 5: Figure extracted from [14] showing the transfer learning task (II) between two MABs. (a) shows a causal graph with known context, U. (b) shows a standard MAB with unobserved confounder (unknown context) indicated with a dotted directed edge.

At this point it is natural to assume that if the identifiability condition does not hold then prior data is not useful in the transfer process. Remarkably though, [14] shows that for non-identifiable tasks we can still obtain causal bounds over the expected rewards of the target agent. This is achieved by using prior knowledge to construct a general SCM compatible with all the available models. Let us consider this is some detail. Given an stochastic MAB problem such that it has a prior represented as a list of bounds over the expected rewards, then for any bandit arm x, let $\mu_x \in [l_x, h_x]$. WLOG, assume $0 < l_x < h_x < 1$ and denote $l_{max} = \max_{x=1,\dots,K} l_x$. Note that a K-MAB problem is simply a generalisation of the MAB problem to multiple independent bandits. The following theorem is presented and proved by the authors:

Theorem 4.1. *Consider a K-MAB problem with rewards bounded in* $[0,1]$*, with each arm* $x \in \{1, \dots, K\}$*, and expected reward* $\mu_x \in [l_x, h_x]$ *s.t.* $0 < l_x < h_x < 1$*. Taking* $f(t) = \log(t) + 3\log(\log(t))$*, in the B-kl-UCB algorithm (shown in algorithm 4.1), the number of draws of* $\mathbb{E}[N_x(T)]$ *for any sub-optimal arm a is upper bounded for any horizon* $T \geq 3$ *as:*

$$\begin{cases} 0 & \text{if } h_x < l_{\max} \\ 4 + 4e\log(\log(T)) & \text{if } h_x \in [l_{\max}, \mu^*) \\ \frac{\log(T)}{KL(\mu_x, \mu^*)} + \mathcal{O}\left(\frac{\log(\log(T))}{KL(\mu_x, \mu^*)}\right) & \text{if } h_x \geq \mu^* \end{cases}$$

Though seemingly very abstract, this theorem tells us that if the causal bounds impose strong constraints over the arm's distribution then the B-kl-UCB algorithm (see below) provides asymptotic improvements over its unbounded counterpart (the kl-UCB algorithm [16], not presented here). This implies that constraints translate into different regret bounds for the MAB agent. We could expect that finding such constraints over our problem bounds would increase performance. The algorithm below should be self-contained. For more information refer to the source material.

Algorithm 1 B-kl-UCB

```
Result: Compute causal bounds for non-identifiable transfer task
Input: Non-decreasing function $f : \mathbb{N} \to \mathbb{R}$
 Input: A list of bounds over $\mu_x : \{[l_x, h_x]\}_{x \in \{1,\dots,K\}}$.
 Remove any arm $a$ with $h_x < l_{max}$. Let $K'$ denote the number of remaining arms. Pull each arm of $\{1, \dots, K\}$
 once. for $t = K^{prime}$ to $T - 1$ do
    for each arm $x$ do
        Compute $\hat{U}_x(t) = \min\{U_x(t), h_x\}$ where $U_x(t) = \sup\{\mu \in [0,1] : KL(\hat{\mu}_x(t), \mu) \leq \frac{f(t)}{N_x(t)}\}$.
    end
    Pick an arm $X_t = \arg\max_{x \in \{1,\dots,K'\}} \hat{U}_x(t)$.
end
```

Zhang and Bareinboim [17] extend similar ideas to the field of dynamic treatment regimes (DTRs) and personalised medicine. A DTR consists of a set of decision rules controlling the provided treatment at any given stage, given a patient's conditions. The challenge is to apply online reinforcement learning algorithms to the problem of selecting optimal DTRs given observational data, with the hope that RLs sample efficiency success in other decision making processes can translate to DTRs. Policy learning in this case refers to the process of finding an optimal policy π that maximises some outcome Y - usually the patient's recovery or improvement in health markers. Often, however, the parameters of the DTR remain unknown and direct optimisation isn't possible. Traditional algorithms rely on there being no unobserved confounders, while randomisation techniques are often not feasible in the medical domain. We certainly wouldn't want doctors randomly testing strategies on patients to see 'how it plays out'! Reinforcement learning offers an attractive set of techniques for DTRs as it should offer an efficient means to learn DTRs while balancing the exploration of state-space and exploitation of rewards. Existing RL techniques, however, are not applicable in the DTR context as they rely on the Markov property. DTRs are clearly non-Markovian as the treatment procedure at some point in the future is a function of past treatments. The authors formalise this in causal language as follows:

Definition 4.1 (Dynamic Treatment Regime [17]). *A dynamic treatment regime (DTR) is a SCM $\langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{u})\rangle$ where the endogenous variables $\boldsymbol{V} = \{\overline{\boldsymbol{X}}_K, \overline{\boldsymbol{S}}_K, Y\}$ are the total stages of interventions. Here $\overline{\boldsymbol{X}}_K$ represents a sequence $\{X_1, \dots, X_K\}$. For stage $k = 1, \dots, K$:*

1. *X_k is a finite decision decided by a behaviour policy $x_k \leftarrow f_k(\overline{\boldsymbol{s}}_k, \overline{\boldsymbol{x}}_{k-1}, \boldsymbol{u})$.*
2. *S_k is a finite state decided by a transition function $s_k \leftarrow \tau_k(\overline{\boldsymbol{x}}_{k-1}, \overline{\boldsymbol{s}}_{k-1}, \boldsymbol{u})$.*
3. *Y is the primary outcome at the final state K, decided by a reward function $y \leftarrow r(\overline{\boldsymbol{x}}_K, \overline{\boldsymbol{s}}_K, \boldsymbol{u})$ bounded in $[0, 1]$.*

Values of exogenous variables $\boldsymbol{U}$ are drawn from the distribution $P(\boldsymbol{u})$.

A DTR M^* induces some observational distribution $P(\overline{\boldsymbol{x}}_K, \overline{\boldsymbol{s}}_K, \boldsymbol{u}))$, responsible for the data we observe without intervention. A policy π for the DTR defines some sequence of stochastic interventions $do(X_1 \sim \pi_1(X_1 \mid \overline{\boldsymbol{S}}_1), \dots, \pi_K(X_K \mid \overline{\boldsymbol{S}}_K, \overline{\boldsymbol{X}}_{K-1}))$. These interventions induce an interventional distribution

$$P_\pi(\overline{\boldsymbol{x}}_K, \overline{\boldsymbol{s}}_K, y) = P_{\overline{\boldsymbol{x}}_K}(y \mid \overline{\boldsymbol{s}}_K) \prod_{k=0}^{K-1} P_{\overline{\boldsymbol{x}}_k}(s_{k+1} \mid \overline{\boldsymbol{s}}_k)\, \pi_{k+1}(x_{k+1} \mid \overline{\boldsymbol{s}}_{k+1}, \overline{\boldsymbol{x}}_k),$$

where $P_{\overline{\boldsymbol{x}}_k}(s_{k+1} \mid \overline{\boldsymbol{s}}_k)$ is the transition distribution at stage k, and $P_{\overline{\boldsymbol{x}}_K}$ is the reward distribution over the primary outcome. The expected cumulative reward is then $V_\pi(M^*) = \mathbb{E}_\pi[Y]$, implying our task is to find $\pi^* = \arg\max_{\pi \in \Pi} V_\pi(M^*)$. In other words, we would like to find the policy that finds the best sequence of actions that leads to an optimal outcome. The notation V_π is deliberately chosen to correspond to value functions in RL literature.

The authors introduce the UC-DTR algorithm, presented in algorithm (4.1) below, to optimise an unknown DTR. This algorithm takes an *optimism in the face of uncertainty* approach - a common strategy in the reinforcement learning literature. Given only knowledge of the state and action domains, UC-DTR achieves near-optimal total regret bounds. This is really quite remarkable. Knowing only about the current state and the possible actions we can take, we have an algorithm to reach almost optimal outcomes in very few steps! We now delve into the algorithm a bit deeper and discuss the overall strategy it employs. First, a new policy π_t is proposed using samples $\{\overline{\boldsymbol{S}}^i_K, \overline{\boldsymbol{X}}^i_K, Y^i\}_{i=1}^{t-1}$ collected up until the current episode, t. That is, the deciding agent exploits its current knowledge to propose what it believes to be a good policy choice. The empirical estimates for the expected reward and the transitional probabilities are calculated and used to consider a set of plausible DTRs in terms of a confidence region around these estimates. The optimal policy of the most optimistic DTR in the plausible DTR set is calculated and executed to collect the next set of samples. This is the *optimism* and the *uncertainty* we discussed earlier. This procedure is repeated until a tolerance level or specific episode is reached. The authors proceed to show that the the UC-DTR algorithm has cumulative regret that scales with $\tilde{\mathcal{O}}(K\sqrt{|\boldsymbol{S}||\boldsymbol{X}|T})$, where $\tilde{\mathcal{O}}(\cdot)$ is like Big-Oh notation but also ignores log-terms. Formally, $f = \tilde{\mathcal{O}}(g) \Leftrightarrow \exists k, f = \mathcal{O}(g \log^k(g))$. The proofs for this analysis are fairly involved and are provided in the appendix of [17].

Theorem 4.2. *Fix tolerance parameter $\delta \in (0, 1)$. With probability at least $1 - \delta$, it holds for any $T > 1$ that the regret of UC-DTR is bounded by*

$$R(T) \leq 12K\sqrt{|\mathcal{S}||\mathcal{X}|T \log(2K|\mathcal{S}||\mathcal{X}|T/\delta)} + 4K\sqrt{T\log(2T/\delta)}.$$

Theorem 4.3. *For any algorithm $\mathcal{A}$, any natural numbers $K \geq 1$, and $|\boldsymbol{\mathcal{S}}^k| \geq 2$, $|\boldsymbol{\mathcal{X}}^k| \geq 2$, for any $k \in \{1, \dots, K\}$, there is a DTR M with horizon K, state domains $\boldsymbol{\mathcal{S}}$ and action domain $\boldsymbol{\mathcal{X}}$, such that the expected regreat of $\mathcal{A}$ after $T \geq |\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|$ episodes ia at least*

$$\mathbb{E}[R(T)] \geq 0.05\sqrt{|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|T}.$$

Together, theorems (4.2) and (4.3) indicate that UC-DTR is near-optimal given only the state and action domains. The authors further propose exploiting available observational data to improve performance of the online learning procedure in the face of unobserved confounders and non-indentifiability. Using observational data, the authors derive theoretically sound bounds on the the system dynamics in DTRs. We include the full UC-DTR algorithm below as it is indicative of a general algorithmic approach to similar problems in the causal reinforcement literature. The reader is encouraged to work through the steps to confirm that it matches the theory discussed above. The explanations for the steps are fairly self-contained and are not discussed further for brevity.

Algorithm 2 UC-DTR

```
Result: Optimise policy for unknown DTR.
Input:  Failure tolerance parameter $\delta \in (0,1)$.
for episodes $t = 1, 2, \dots$ do
    Define event counts $N^t(\overline{\boldsymbol{s}}_k, \overline{\boldsymbol{x}}_k) = \sum_{i=1}^{t-1} I_{\overline{\boldsymbol{S}}^i_k = \overline{\boldsymbol{s}}_k, \overline{\boldsymbol{X}}^i_k = \overline{\boldsymbol{x}}_k}$ and $N^t(\overline{\boldsymbol{s}}_k, \overline{\boldsymbol{x}}_{k-1}) = \sum_{i=1}^{t-1} I_{\overline{\boldsymbol{S}}^i_k = \overline{\boldsymbol{s}}_k, \overline{\boldsymbol{X}}^i_{k-1} = \overline{\boldsymbol{x}}_{k-1}}$ for horizon $k = 1, \dots, K$.
    Define reward count $R^t(\overline{\boldsymbol{s}}_K, \overline{\boldsymbol{x}}_K) = \sum_{i=1}^{t-1} Y^i I_{\overline{\boldsymbol{S}}^i_K = \overline{\boldsymbol{s}}_K, \overline{\boldsymbol{X}}^i_K = \overline{\boldsymbol{x}}_K}$. Compute estimates
        $\hat{P}^t_{\overline{\boldsymbol{x}}_k}(s_{k+1} \mid \overline{\boldsymbol{s}}_k) = \frac{N^t(\overline{\boldsymbol{s}}_{k+1}, \overline{\boldsymbol{x}}_k)}{\max\{1, N^t(\overline{\boldsymbol{s}}_k, \overline{\boldsymbol{x}}_k)\}}, \quad \hat{E}^t_{\overline{\boldsymbol{x}}_K}[Y \mid \overline{\boldsymbol{s}}_K] = \frac{R^t(\overline{\boldsymbol{s}}_K, \overline{\boldsymbol{x}}_K)}{\max\{1, N^t(\overline{\boldsymbol{s}}_k, \overline{\boldsymbol{x}}_k)\}}.$
    Let $\mathcal{M}_t$ denote a set of DTRs such that for any $M \in \mathcal{M}_t$ its transition probabilities and rewards are close estimates. Formally,
        $\left\| P_{\overline{\boldsymbol{x}}_k}(\cdot \mid \bar{s}_k) - \hat{P}^t_{\overline{\boldsymbol{x}}_k}(\cdot \mid \overline{\boldsymbol{s}}_k) \right\|_1 \le \sqrt{\frac{6|\mathcal{S}_{k+1}|\log(2K|\overline{\boldsymbol{\mathcal{S}}}_k||\overline{\boldsymbol{\mathcal{X}}}_k|t/\delta)}{\max\{1, N^t(\overline{\boldsymbol{s}}_k, \overline{\boldsymbol{x}}_k)\}}}$
        $\left| E_{\overline{\boldsymbol{x}}_K}[Y \mid \overline{\boldsymbol{s}}_K] - \hat{E}^t_{\overline{\boldsymbol{x}}_K}[Y \mid \overline{\boldsymbol{s}}_K] \right| \le \sqrt{\frac{2\log(2K|\boldsymbol{\mathcal{S}}||\boldsymbol{\mathcal{X}}|t/\delta)}{\max\{1, N^t(\bar{s}_K, \overline{\boldsymbol{x}}_K)\}}}$
    Find optimal policy $\pi_t$ of optimistic DTR $M_t \in \mathcal{M}_t$ such that
        $V_{\boldsymbol{\pi}_t}(M_t) = \max_{\boldsymbol{\pi} \in \boldsymbol{\Pi}, M \in \mathcal{M}_t} V_{\boldsymbol{\pi}}(M).$
    Execute policy $\pi_t$ for episode $t$ and observe the samples $\overline{\boldsymbol{S}}^t_K, \overline{\boldsymbol{X}}^t_K, Y^t$.
end
```

We have discussed some interesting theory that underlines much of the later work in this active area of research. The interested reader is encouraged to refer to [18] for extensions to dynamic treatment regimes. [19] and [20] will interest the readers motivated by the development of further theory for generalised decision making and performance bounding. We are now ready to discuss the problem of finding where in a causal system we should intervene for optimal and efficient outcomes.

4.2 Task 2: Interventions: Where and When?

A classic problem in reinforcement learning literature regards the trade-off between exploration of the state-action space, to test interventional outcomes over a long-term horizon, and weigh this against greedy exploitation of current knowledge of the behaviour of the underlying causal system. The multi-armed bandit (MAB) problem is a classic setting for the study of decision making agents - a situation where long-term planning is not required. Recent literature has placed focus on the effect of non-trivial dependencies amongst arms of the bandits, referred to as *structural bandits*. Researchers interested in causal reasoning have experimented and investigated modelling some of these dependencies with causal graph structures. For example, [21] discusses reinforcement learning in the context of causal models with unobserved confounders. This is discussed in detail in section 4.3 below, with key insight being that counterfactual quantities can aid in taking into account unobserved confounders. In this setting, traditional methods do not guarantee convergence to a reasonable policy while counterfactual-aware methods can make this guarantee. The goal of this section is to identify the optimal action an MAB should take where the arm of the MAB corresponds to an intervention on some causal graph. In this way we hope to use knowledge of causal systems to identify theoretically optimal interventional decisions. To discuss MAB instances in the context of causal inference, we now formalise the notion of an SCM-MAB.

Definition 4.2 (Structural Causal Model - Multi-Armed Bandit (SCM-MAB) [22])**.** *Let M be a SCM $\langle \boldsymbol{U}, \boldsymbol{V}, \boldsymbol{F}, P(\boldsymbol{U}) \rangle$ and $Y \in \boldsymbol{V}$ be a reward variable with domain $\mathbb{R}$. The arms of the bandit are such that $\{\boldsymbol{x} \in Dom(\boldsymbol{X}) \mid \boldsymbol{X} \subseteq \boldsymbol{V} \setminus \{Y\}\}$. In other words, the arms are the set of possible interventions on all the endogenous variables of the SCM, excluding the reward. Each arm of the MAB is associated with a reward interventional distribution $P(Y \mid do(x))$, having mean $\mu_x = \mathbb{E}[Y \mid do(x)]$.*

We assume the SCM-MAB has knowledge of the underlying causal graph and reward, but does not know about the SCM mappings $\boldsymbol{F}$, or the joint distribution over the exogenous variables $P(\boldsymbol{U})$. Recall, we are interested in finding a minimal set of interventions that optimises the actions. This motivates the following definitions. Here, we denote the information gained by an agent interacting with the SCM-MAB by $[\![G, Y]\!]$.

Definition 4.3 (Minimal Intervention Set (MIS) [22])**.** *A set of endogenous variables $\boldsymbol{X} \subseteq \boldsymbol{V} \setminus \{Y\}$ is said to be a minimal intervention set relative to $[\![G, Y]\!]$ if $\nexists$ $\boldsymbol{X}' \subset \boldsymbol{X}$ such that $\mu_{x'} = \mu_x$ where $x' \in \boldsymbol{X}'$ for every SCM complying with the causal structure dictated by the causal graph. In other words, there is no intervention set smaller than the proposed set that results in the same mean reward.*

Thinking about the definition, we realise that by intervening on the ancestors of the reward variable, Y, is a sufficient and necessary condition for a MIS.

Proposition 4.1 (Minimality [22])**.** *A set of endogenous variables $\boldsymbol{X} \subseteq \boldsymbol{V} \setminus \{Y\}$ is a MIS for the causal graph G and reward variable Y if and only if $\boldsymbol{X} \subseteq an(Y)_{G_{\overline{X}}}$.*

This proposition provides a method for finding the MISs given the information available to the agent. We can simply pass over all the possible subsets of the endogenous variables $\boldsymbol{X} \setminus \{Y\}$ and check the proposition. Of course, based on the structure of causal graphs, it is possible that intervening on one set of variables is always superior to intervening on another. This induces a partial ordering: given two sets of variables $\boldsymbol{W}, \boldsymbol{Z} \subseteq \boldsymbol{V} \setminus \{Y\}$ such that

$$\max_{\boldsymbol{w} \in Dom(\boldsymbol{W})} \mu_{\boldsymbol{w}} \leq \max_{\boldsymbol{z} \in Dom(\boldsymbol{Z})} \mu_{\boldsymbol{z}}$$

in all the possible SCMs that comply with the rules defined by the causal graph, then we have a method of comparing the preferable intervention sets. This motivates the need to formalise the notion of a possibly optimal MIS.

Definition 4.4 (Possibly-Optimal MIS (POMIS) [22])**.** *Given information $[\![G, Y]\!]$ and MIS $\boldsymbol{X}$. If there exists a SCM following rules defined by causal graph G and $\mu_{\boldsymbol{x}^*} > \mu_{\boldsymbol{z}^*} \forall \boldsymbol{Z} \in \mathbb{Z} \setminus \{\boldsymbol{X}\}$, we say $\boldsymbol{X}$ is a possibly optimal MIS. Here $\mathbb{Z}$ represents the possible MISs complying with G and Y.*

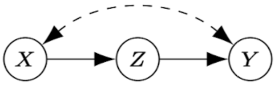


Figure 6: Figure showing basic example of causal model with unobserved confounder affecting X and Y. Here intervening on Z is possibly optimal. However, if intervening on Z is not feasible, intervening on X should be regarded as the possibly optimal solution (along with no intervention). This motivates the extension of the POMIS definition (4.4). Figure extracted from [17].

One key assumption we have made is that any observed variable can be intervened on. This is, of course, a ludicrous assumption to be made in general, and the studying of the relaxed condition becomes relevant for general application of these techniques. [23] investigates identifying possibly-optimal arms in the structural causal bandit framework when only partial causal knowledge is available and select variables are not manipulable. See the caption in figure (6) and notice that according to the POMIS definition, intervening on Z is possibly optimal. If intervening on this variable is not feasible (i.e. non-manipulable) then we should consider X to be possibly optimal instead. Of course, we could have that not intervening at all ($do(\theta)$) is optimal. We can augment the definitions of MIS and POMIS by replacing the endogenous variables under consideration, $\boldsymbol{X} \subseteq \boldsymbol{V} \setminus \{Y\}$ with $\boldsymbol{X} \subseteq \boldsymbol{V} \setminus \{Y\} \setminus \boldsymbol{N}$, where $\boldsymbol{N}$ represents the variables that are non-manipulable in the causal system. For brevity we do not formalise this as it follows directly by augmenting previous definitions.

Identifying all the POMIS given that we have $\boldsymbol{N} \neq \theta$ (i.e. there are non-manipulable variables) is non-trivial because simply removing intervention sets that contain manipulable variables will not necessarily be exhaustive of the possibly-optimal sets. If this is not clear, refer back to figure (6) and notice that the possibly-optimal set $\{X\}$ (when Z is non-manipulable) would not be identified in this case. This motivates the need for new techniques. The root of the problem is that there is a possibility for a POMIS to exist that would not be a POMIS in the unconstrained case. The idea presented in [17] is to project the constrained graph onto a subgraph containing only unconstrained variables, and then identify the (unconstrained) POMIS structures in this realm. We now describe the method [17] presents to project a constrained causal graph onto the unconstrained graph, ready for POMIS identification.

1. Initialise graph $H = \langle \boldsymbol{V} \setminus \boldsymbol{N}, \theta \rangle$.
2. Add edge $V_i \rightarrow V_j$ if there exists path $V_i \rightarrow V_j$ in the original causal graph G; or
3. Add edge $V_i \rightarrow V_j$ if there exists a directed path from V_i to V_j where all intermediate nodes are in $\boldsymbol{N}$.
4. Similarly, add edge $V_i \leftrightarrow V_j$ if there exists path $V_i \leftrightarrow V_j$ in the original causal graph; or

5. Add edge $V_i \leftrightarrow V_j$ if there exists a bidirected (unobserved confounder) path from V_i to V_j where all intermediate nodes are in $\boldsymbol{N}$.

Figure 7 displays some examples of this projection procedure for removing different variables. The authors proceed to show that this projection is guaranteed to preserve the POMISs, and that the underlying SCMs of both graphs (original and projected) maintain the same observational and interventional distributions. By applying these techniques to simulated datasets, the authors show that the POMIS technique significantly outperforms brute-force intervention techniques and simpler MIS identification to identify optimal arm selection. They also experiment with adding z-identifiability (discussed in later sections) to the POMIS method, which improves it further.

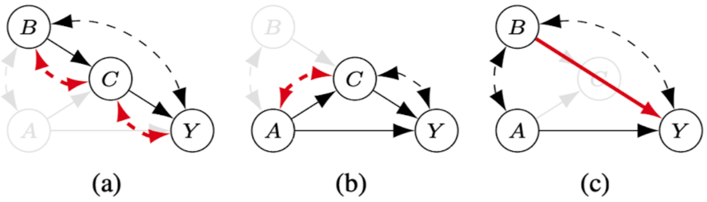


Figure 7: Figure showing examples of projected graphs from marginalising out A, B, and C respectively, by following the projection procedure outlined. Figure extracted from [17].

In this section we have briefly discussed some graphical criteria and procedures for identifying optimal sets of variables to intervene on. [24] extend these ideas to identifying optimal intervention variables in mixed policies - cases where both observational and interventional information is available. The reader is encouraged to refer to this paper. One key idea presented in work in this area is that of z-identifiability which encapsulates the idea of what we can learn by intervening on variables in the causal system. This idea is generalised in later sections (see task 4). This leads well to the next section which discusses a key theme in this paper - counterfactual decision making.

4.3 Task 3: Counterfactual Decision Making

A key feature of causal inference is its ability to deal with counterfactual queries. Reinforcement learning, by its nature, deals with interventional quantities in a trial-and-error style of learning. Perhaps the most obvious question is: how can we implement RL in such a way as to deal with counterfactual and other causal factors in uncertain environments? In the preliminaries section we discussed the notions of a multi-armed bandit and their associated policy regret. This is a natural starting point for the merger of reinforcement learning and causal inference theory to solving counterfactual decision problems. We now discuss some interesting work done in this area, especially in the context of unobserved confounders in decision frameworks.

Obviously, the goal of optimising a policy is to minimise the regret an agent experiences. In addition, it should achieve this minimum regret as soon as possible in the learning process. Auer et al. [25] show that policies can achieve logarithmic regret uniformly over time. Applying a policy they call *UCB2*, the authors show that with input $0 < \alpha < 1$ run on K bandits (think slot machines), the expected regret after any number $n > \max_{i:\mu_i<\mu^*} \frac{1}{2\Delta_i^2}$ of plays is bounded by

$$\sum_{i:\mu_i<\mu^*}\left(\frac{(1+\alpha)(1+4\alpha\ln(2e\Delta_i^2 n))}{2\Delta_i}+\frac{c_\alpha}{\Delta_i}\right),$$

where $\Delta_i = \mu_i^* - \mu_i$. Under different assumptions similar bounds can be achieved. Bareinboim et al. [21] show that these strategies are complicated by unobserved confounders when they are present in the data generating process. In fact, previous bandit algorithms implicitly attempt to maximise rewards by estimating experimental distributions. In other words, they attempt to optimise the procedure by calculating how their actions influence outcomes. This strategy fails to guarantee an optimal strategy in the wake of confounders - that is, unobserved factors in the system. Rephrasing the multi-armed bandit problem to account for unobserved confounders in causal language leads to a principle which exploits both observational and experimental modes of data to optimise reward. We now discuss this formalism in some detail with a motivating example in mind.

The following example is adapted from [21]. Imagine a casino with slot machines designed to detect whether a gambler is drunk. These machines are designed in such a way that they can attract drunk gamblers by flashing a light (stimulating some interest in drunk gamblers), knowing full well that drunk gamblers are less likely to notice machines tailoring payouts such that they are exploited. With full knowledge of gambling law, the casino devises a strategy to fool random testing strategies such that it appears the casino is following the legally required 30% minimum payout. As enlightened scholars aware of causal inference theory, we have knowledge of the causal structure of the payout scheme and decide to condition on the intent of the gamblers. That is, we stratify data according to which machine a gambler *intends* to play, realising payout and intent are confounded by sobriety. The idea of an actor's intent can be formalised. This will be very useful throughout CRL theory development.

Definition 4.5 (Intent [26])**.** *For all variables requiring an actor's decision $\Pi_i \in \Pi$ in an SCM M, let the actor's intended choice $I_{\Pi_i,t} = i_{\Pi_i,t}$ be the choice that the actor would make observationally for unit t and the present unit's configuration of UCs $U_t = u_t$. Formally, for parents of Π_i, $pa(\Pi_i)$, let $I_{\Pi_i,t} = f_{\Pi_i}(pa(\Pi_i)_t, u_{\Pi_i,t})$.*

Applying the idea of conditioning on intent reveals gamblers are actually being paid 15% (see table 1).

(a)	D = 0		D = 1	
	B=0	B=1	B=0	B=1
$X = M_1$	0.10*	0.50	0.40	0.20*
$X = M_2$	0.50	0.10*	0.20*	0.40

(b)	$p(y \mid X)$	$p(y \mid do(X))$
$X = M_1$	0.15	0.3
$X = M_2$	0.15	0.3

Table 1: Table (a) encodes slot machine payout probabilities as found in the casino as a function of sobriety D, whether or not the machine has a flashing light B, and the machine type X. The 'natural' choices - or intent - of the gamblers are indicated by asterisks. Table (b) shows the observational and interventional probabilities of payouts when naively intervening on machines. Naive randomisation in the face of unobserved confounders (blinking light) fails to reveal the violation of gambling law. Tables recreated from [21].

Definition 4.6 (K-Armed Bandit with Unobserved Confounders [21])**.** *A K-Armed bandit problem with unobserved confounders is a causal model M with reward distribution over $P(u)$ where:*

1. *$X_t \in \{x_1, \dots, x_n\}$ is an observable encoding an agent's arm choice from one of k arms. This is decided naturally in the observational case, and by the policy in the experimental case, $do(X_t = \pi(x_0, y_0, \dots, x_{t-1}, y_{t-1}))$ for strategy π.*
2. *U_t represents the unobserved variable which affects the payout rate of arm x_t by nature of influencing the agent's choice.*
3. *$Y_t \in \{0, 1\}$ is a reward (0 indicates loss, 1 indicates a win) for choosing arm x_t. This reward is determined by $y_t = f_y(x_t, u_t)$.*

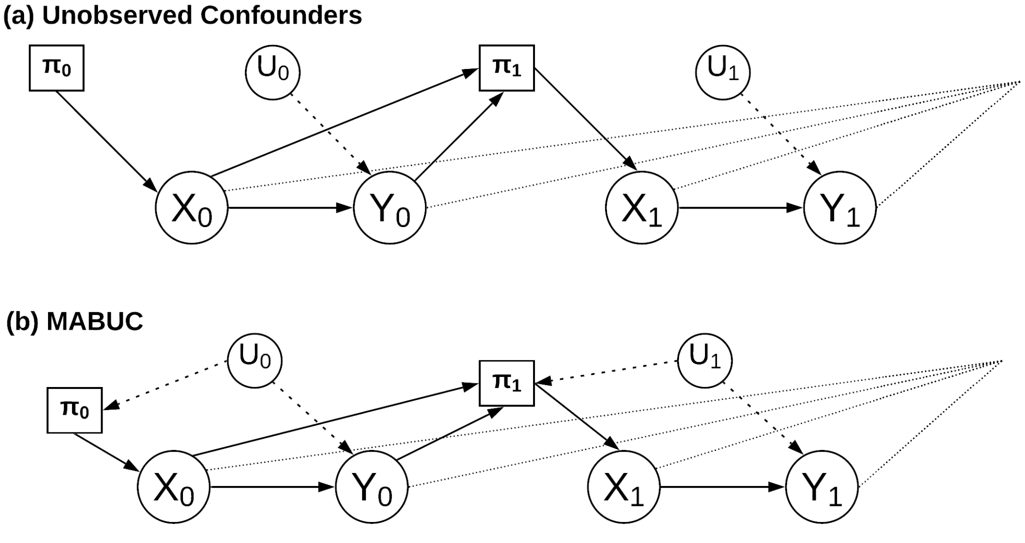


Figure 8: Diagram showing how unobserved confounder affects the MAB decision making process. (a) shows the standard MAB procedure in which only previous states and outcomes influence the policy. (b) shows the MABUC case in which unobserved confounders directly affect the policy of the agent. This changes how and what an agent can learn and is what [21] addresses. Figure created for this paper.

Allowing gamblers free roam of the casino floor to play slot machines as desired corresponds to the natural policy, π_t. This policy induces the observational distribution $P(y \mid x)$. Randomisation, on the other hand, removes external policies and corresponds to the interventional distribution $P(y \mid do(x))$. A major contribution of [21] is the extraction of information contained in both data-collection (observational and interventional) modes to understand and account for the players' natural instincts. To perform such an analysis we need to consider counterfactuals: "would the agent win had it played on machine M_i instead of M_j?" In other words, instead of considering $\arg\max_a \mathbb{E}[Y \mid do(X = a)]$ we should should be considering

$$\arg\max_{a} \mathbb{E}\left[Y_{X=a} = 1 \mid X = x\right],$$

where x represents the players intent and a represents their final decision. This forms the *regret decision criterion* (RDC) - maximise the expected reward condition on the agents intent - and allows the following heuristic:

$$\mathbb{E}(Y_{X=0} = 1 \mid X = 1) > \mathbb{E}(Y_{X=1} = 1 \mid X = 1) \Leftrightarrow \mathbb{E}(Y_{X=0} = 1 \mid X = 1) > P(Y \mid X = 1).$$

All this says is that we should consider the intent of the agent and consider how the causal system is trying to exploit this intent. The agent should acknowledge it's own intent and modify its policy accordingly. The authors propose incorporating this into Thompson Sampling [27] to form Causal Thompson Sampling (TS^C), shown in algorithm (4.3) below. This algorithm leverages observational data to seed the algorithm and adds a rule to improve the arm exploration in the MABUC case. The algorithm is shown to dramatically improve convergence and regret against traditional methods under certain scenarios. Comments have been added to make the algorithm self-contained. The reader is encouraged to work through it.

Algorithm 3 Causal Thompson Sampling (TS^C) [21].

```
Result: Optimise MABUC policy by considering intent
$\mathbb{E}[Y_{X=a} \mid X] \leftarrow P_{obs}(y \mid X)$                                   (Seed)
for $t = [1, \dots, T]$ do
    $x \leftarrow intuition(x)$                                                  (Intuition)
    $Q_1 \leftarrow \mathbb{E}[Y_{X=x'|X=x}]$                                          (Counter-intuition)
    $Q_2 \leftarrow P(y \mid X = x)$                                           (Estimate reward)
    $w \leftarrow [1, 1]$                                                      (Initialise weights)
    $bias \leftarrow 1 - |Q_1 - Q_2|$                                       (Compute weighting strength)
    if $Q_1 > Q_2$ then $w[x] \leftarrow bias$ else $w[x'] \leftarrow bias$                  (Choose arm to bias)
    $a \leftarrow \max(\beta(s_{M_1,x}, f_{M_1,x}) \times w[1], \beta(s_{M_2,x}, f_{M_2,x}) \times w[2])$   (Choose arm)
    $y \leftarrow pull(a)$                                                       (Reward)
    $\mathbb{E}[Y_{X=a} \mid X = x] \leftarrow y|a, x$                                    (Update)
end
```

The TS^C is empirically shown to outperform conventional methods by converging upon an optimal policy faster and with less regret than the non-causally empowered approach. Ultimately we are interested in sequential decision making in the context of planning - the cases reinforcement learning addresses. [28] generalises the previous work on MABs to take a causal approach to Markov decision processes (MDPs) with unobserved confounders. In a similar fashion to [21], the authors construct a motivating example and show that MDP algorithms perform sub-optimally when applying conventional (non-causal) methods. First, we note why confounding in MDPs is different to confounding in MABs. Confounding in MDPs affects state and outcome variables in addition to action and outcome, and thus requires special treatment. Unlike in MABs, the MDP setting requires maximisation of reward over a long-term horizon (planning). Maximising with respect to the immediate future (greedy behaviour) thus fails to account for potentially superior long term trajectories in state-action space (long term strategies). The authors proceed by showing conventional MDP algorithms are not guaranteed to learn an optimal policy in the presence of unobserved confounders. Reformulating MDPs in terms of causal inference, we can show that counterfactual-aware policies outperform purely experimental algorithms.

Definition 4.7 (MDP with Unobserved Confounders (MDPUC) [28])**.** *A Markov Decision Process with Unobserved Confounders is an SCM M with actions X, states S, and a binary reward Y:*

1. $\gamma \in [0, 1)$ *is the discount factor.*
2. $U^{(t)} \in U$ *is the unobserved confounder at time-step* t*.*
3. $V^{(t)} = X^{(t)} \cup Y^{(t)} \cup S^{(t)}$ *is the set of observed variables at time-step* t*, where* $X^{(t)} \in X$*,* $Y^{(t)} \in Y$*, and* $S^{(t)} \in S$*.*
4. $F = \{f_x, f_y, f_s\}$ *are the set of structural equations relative to* V *such that* $X^t \leftarrow f_x(s^{(t)}, u^{(t)})$*,* $Y^{(t)} \leftarrow f_y(x^{(t)}, s^{(t)}, u^{(t)})$*, and* $S^{(t)} \leftarrow f_s(x^{(t-1)}, s^{(t-1)}, u^{(t-1)})$*. In other words, they determine the next state and associated reward.*
5. $P(u)$ *encodes the probability distribution over the unobserved (exogenous) variables* U*.*

A key difference to the MABUC case is that different sets of variables can be confounded over the time horizon. The following figure shows the four different ways in which MDPUC variables can be confounded.

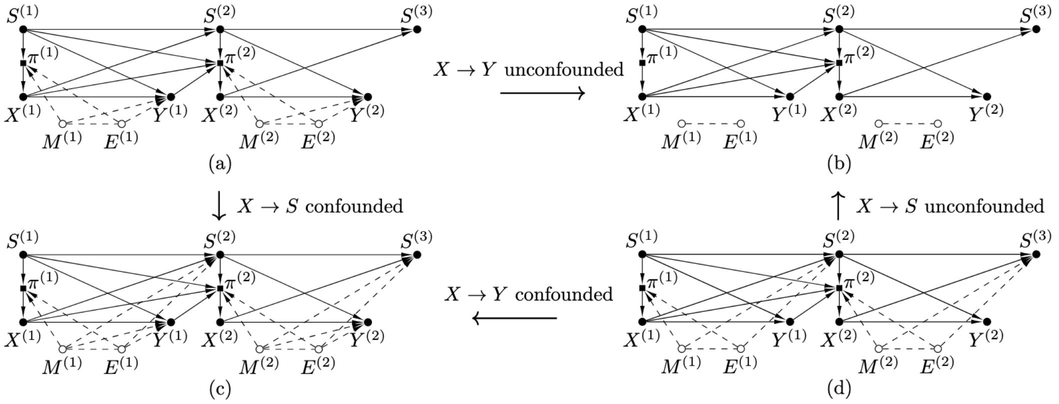


Figure 9: (a) MDPUC with action (decided by the policy) to reward path, $x^{(t)} \to y^{(t)}$ confounded. (b) MDP without confounders. This is the traditional RL instance. (c) MDPUC with action to reward path, $x^{(t)} \to y^{(t)}$, and action to state path, $x^{(t)} \to s^{(t+1)}$, confounded. (d) MDPUC with only action to state path, $x^{(t)} \to s^{(t+1)}$, confounded. Extracted from [28].

Most literature discussing reinforcement learning will develop the theory in terms of value functions. We develop these notions for the MDPUC case here. This will allow us (in theory) to exploit existing RL literature with relative ease.

Definition 4.8 (Value Functions). *Given a MDPUC model $M\langle\gamma, U, X, Y, S, F, P(u)\rangle$ and an arbitrary deterministic policy π, we can define the value function starting from state $s^{(t)}$, taking action $x^{(t)}$, and thereafter following policy π as:*

$$V^{\pi}(s^{(t)}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k Y^{(t+k)}_{x([t,t+k])=\pi} \mid s^{(t)}\right].$$

The state-action value function is similarly defined:

$$Q^{\pi}(s^{(t)}, x^{(t)}) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k Y^{(t+k)}_{x^{(k)}, x([t+1,t+k])=\pi} \mid s^{(t)}, x^{(t)}\right].$$

The interpretation of these functions is the same as in the usual RL literature. They simply associate a state or state-action with an expected reward over all future time steps. Using these definitions, we can derive expressions for the well known Bellman equation and recursive definitions for the state and state-action value functions [12] for the situation where unobserved confounders are present. The authors rest their analysis of these results on the axioms of counterfactuals and the Markov property presented in theorem (4.4). Proofs for the following theorems are available in the source paper [28] and are not included here for brevity.

Theorem 4.4 (Markovian Property in MDPUCs [28]). *For a MDPUC model $M\ \langle\gamma, U, X, Y, S, F, P(u)\rangle$, a policy $\pi \in F_{exp} = \{\pi \mid \pi : S \to X\}$ (state to action map), and a starting state $s^{(t)}$, the agents performs actions $do(X^{(t)} = x^{(t)})$ at round t and $do(X^{([t+1,t+k])} = \pi)$ afterwards $(k \in \mathbb{Z}^+)$, the following statement holds:*

$$P\left(Y^{t+k}_{x^{(t)}, x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}_{x^{(t)}}, s^{(t)}\right) = P\left(Y^{t+k}_{x^{([t+1,t+k])}=\pi} = y^{(t+k)} \mid s^{(t+1)}\right)$$

We can further extend MDPUCs to counterfactual policies by considering $F_{ctf} = \{\pi \mid \pi : S \times X \to X\}$, the set of functions between the current state $s^{(t)}$, the intuition of the agent $x'^{(t)}$, and the action $x^{(t)}$. Then $V^{(t)} = V^{(t)}(s^{(t)}, x'^{(t)})$ and $Q^{(t)} = Q^{(t)}(s^{(t)}, x'^{(t)}, x^{(t)})$. With this we can derive a remarkable result encoded as theorem (4.5) below.

Theorem 4.5. *Given an MDPUC instance $M\langle\gamma, U, X, Y, S, F, P(u)\rangle$, let $\pi^*_{exp} = \arg\max_{\pi \in F_{exp}} V^{\pi}(s^{(t)})$ and $\pi^*_{ctf} = \arg\max_{\pi \in F_{ctf}} V^{\pi}(s^{(t)}, x'^{(t)})$. For any state $s^{(t)}$, the following statement holds:*

$$V^{\pi^*_{exp}}(s^{(t)}) \leq V^{\pi^*_{ctf}}(s^{(t)}).$$

In other words, we can never do worse *by considering counterfactual quantities (intent).*

Zhang and Bareinboim [28] continue to implement a counterfactual-aware MORMAX MDP algorithm and empirically show it superior to conventional MORMAX [29] approaches in intent-sensitive scenarios. Forney and Bareinboim [11] extend this counterfactual awareness to the design of experiments. Forney, Pearl and

Bareinboim [30] expand upon the ideas presented above by showing that counterfactual-based decision-making circumvents problems of naive randomisation when UCs are present. The formalism presented coherently fuses observational and experimental data to make well informed decisions by estimating counterfactual quantities empirically. More concretely, they study conditions under which data collected from distinct sources and conditions can be combined to improve learning performance by an RL agent. The key insight here is that *seeing* does not equate to *doing* in the world of data. Applying a developed heuristic, a variant of Thompson Sampling [27] is introduced and empirically shown to outperform previous state-of-the-art agents. An extension of the motivating example of (potentially) drunk gamblers from [21] is extended to consider an extra dependence on whether or not all the machines have blinking lights. This yields four combinations of the states of sobriety and blinking machines.

We start by noticing the interventional quantities, $\mathbb{E}[Y \mid do(X=x)]$ can be written in terms of counterfactuals as $\mathbb{E}[Y_{X=x}] = \mathbb{E}[Y_x]$ - that is, the expected value of Y had X been x. Applying the law of total probabilities, we arrive at the useful representation

$$\mathbb{E}[Y_x] = \mathbb{E}[Y_x \mid x_1]P(x_1) + \dots + \mathbb{E}[Y_x \mid x_K]P(x_K). \tag{1}$$

$\mathbb{E}[Y_x]$ is interventional by definition. $\mathbb{E}[Y_x \mid x']$ is either observational or counterfactual depending on whether $x = x'$ or $x \neq x'$ respectively. That is, whether or not x' has occurred. As we noted earlier, counterfactual quantities are, by their nature, not empirically available in general. Interestingly, however, intents of the agent contain information about the agent's decision process and can reveal encoded information about unobserved confounders, as we noted in the MAB case earlier. Applying randomisation to intent conditions (with RDC) can allow computation of counterfactual quantities. In this way, observational data actually *adds* information to the seemingly more informative interventional data. This counterfactual quantity - that is not naturally realisable - is often referred to as the *effect of the treatment on the treated* [31], by way of the fact that doctors cannot retroactively observe how changing the treatment would have affected a specific patient - well, they shouldn't!

Theorem 4.6 (Estimation of ETT [30]). *The counterfactual quantity referred to as the effect of the treatment on the treated (ETT) is empirically estimable for any number of action choices when agents condition on their intent $I = i$ and estimate the response Y to their final action choice $X = a$.*

Proof. The ETT counterfactual quantity can be written as $\mathbb{E}[Y_{X=a} \mid X = i]$. Applying law of total probability and conditional independence relation $Y_x \perp\!\!\!\perp X \mid I$, we have:

$$\begin{aligned}
\mathbb{E}[Y_{X=a} \mid X = i] &= \sum_{i'} \mathbb{E}[Y_{X=a} \mid X = i, I = i']P(I = i' \mid X = i) \\
&= \sum_{i'} \mathbb{E}[Y_{X=a} \mid I = i']P(I = i' \mid X = i)
\end{aligned}$$

Now, we notice that $I_x = I$ because in $G_{\overline{X}}$ (graph with edges into X removed) we have $(I \perp\!\!\!\perp X)_{G_{\overline{X}}}$. Thus,

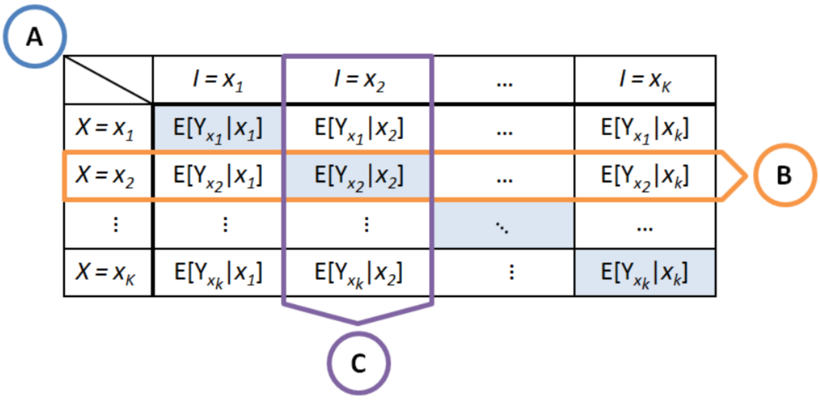
$$\begin{aligned}
\mathbb{E}[Y_{X=a} \mid X = i] &= \sum_{i'} \mathbb{E}[Y_{X=a} \mid I = i']P(I = i' \mid X = i) \\
&= \sum_{i'} \mathbb{E}[Y_{X=a} \mid I_{x=a} = i']P(I = i' \mid X = i) \\
&= \sum_{i'} \mathbb{E}[Y \mid do(X = a), I = i']P(I = i' \mid X = i)
\end{aligned}$$

where the last line follows since all quantities are in relation to the same variable x and so the counterfactual quantity can be written as an interventional quantity. We now notice that since $P(I = i' \mid X = i)$ is observational, the intent will always match the outcome. We can thus rewrite this as an indicator function. The result follows.

$$\begin{aligned}
\mathbb{E}[Y_{X=a} \mid X = i] &= \sum_{i'} \mathbb{E}[Y \mid do(X = a), I = i']P(I = i' \mid X = i) \\
&= \sum_{i'} \mathbb{E}[Y \mid do(X = a), I = i']\mathbb{1}_{i'=i} \\
&= \mathbb{E}[Y \mid do(X = a), I = i]
\end{aligned}$$

□

[30] makes use of this result to suggest heuristics for learning counterfactual quantities from (possibly noisy) experimental and observational data.



	$I = x_1$	$I = x_2$	...	$I = x_K$
$X = x_1$	$E[Y_{x_1} \mid x_1]$	$E[Y_{x_1} \mid x_2]$	...	$E[Y_{x_1} \mid x_k]$
$X = x_2$	$E[Y_{x_2} \mid x_1]$	$E[Y_{x_2} \mid x_2]$	...	$E[Y_{x_2} \mid x_k]$
⋮	⋮	⋮	⋱	...
$X = x_K$	$E[Y_{x_k} \mid x_1]$	$E[Y_{x_k} \mid x_2]$	⋮	$E[Y_{x_k} \mid x_k]$

Figure 10: Figure showing the counterfactual possibilities for different actions and action-intents. The diagonal indicates the counterfactual quantities that have occurred. We can apply knowledge of other counterfactual quantities to learn about other possible counterfactuals. (B) indicates cross-intent learning. (C) indicates cross-arm learning. Extracted form [30].

1. **Cross-Intent Learning:** Thinking about equation (1) carefully, we can notice that since this holds for every arm, we have a system of equations for outcomes conditioned on different intents. Thus to find $\mathbb{E}[Y_{x_r} \mid x_w]$ we can learn about *other* intent conditions:

$$\mathbb{E}[Y_{x_r} \mid x_w] = \left[\mathbb{E}[Y_{x_r}] - \sum_{i \neq w}^{K} \mathbb{E}[Y_{x_r} \mid x_i] P(x_i)\right] / P(x_w).$$

2. **Cross-Arm Learning:** Similarly to (1), we can observe that given information about two different arms under the same intent, we can learn about a third arm under the same intent. We have

$$\begin{aligned} P(x_w) &= \frac{E[Y_{x_r}] - \sum_{i \neq w}^{K} E[Y_{x_r} \mid x_i] P(x_i)}{E[Y_{x_r} \mid x_w]} \\ &= \frac{E[Y_{x_s}] - \sum_{i \neq w}^{K} E[Y_{x_s} \mid x_i] P(x_i)}{E[Y_{x_s} \mid x_w]} \end{aligned}$$

Combining these results and solving for $\mathbb{E}[Y_{x_r} \mid x_w]$ we find:

$$E[Y_{x_r} \mid x_w] = \frac{\left[E[Y_{x_r}] - \sum_{i \neq w}^{K} E[Y_{x_r} \mid x_i] P(x_i)\right] E[Y_{x_s} \mid x_w]}{E[Y_{x_s}] - \sum_{i \neq w}^{K} E[Y_{x_s} \mid x_i] P(x_i)}. \quad (2)$$

This estimate is not robust to noise in the samples. We can take a pooling approach and account for variance of reward payouts by applying an inverse-variance-weighted average:

$$E_{XArm}[Y_{x_r} \mid x_w] = \frac{\sum_{i \neq r}^{K} h_{XArm}(x_r, x_w, x_i) / \sigma^2_{x_i, x_w}}{\sum_{i \neq r}^{K} 1/\sigma^2_{x_i, x_w}},$$

where $h_{XArm}(x_r, x_w, x_s)$ evaluates equation (2), and $\sigma^2_{x_i,i}$ is the reward variance for arm x under intent i.

3. **Combined Approach:** Of course, we can combine the previous two approaches by sampling (collecting) estimates during execution and applying *cross-intent* and *cross-arm* learning strategies. A fairly straightforward derivation yields a combined approach formula:

$$E_{\text{combo}}[Y_{x_r} \mid x_w] = \frac{\alpha}{\beta}, \quad \text{where}$$

$$\begin{aligned} &\alpha = E_{\text{samp}}[Y_{x_r} \mid x_w] / \sigma^2_{x_r, x_w} + E_{\text{XInt}}[Y_{x_r} \mid x_w] / \sigma^2_{\text{XInt}} + E_{\text{XArm}}[Y_{x_r} \mid x_w] / \sigma^2_{XArm} \\ &\beta = 1/\sigma^2_{x_r, x_w} + 1/\sigma^2_{\text{XInt}} + 1/\sigma^2_{XArm} \end{aligned}$$

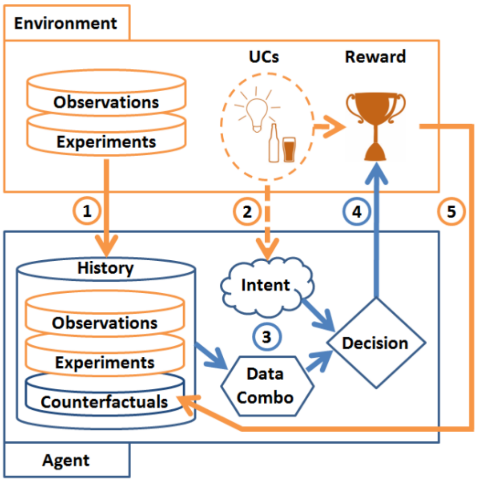


Figure 11: Figure showing the flow and process of fusing data using counterfactual reasoning as outlined in this section. The agent employs both the history of interventional and observational data to compute counterfactual quantities. Along with its intended action, the agent makes a counterfactual and intent aware decision to account for unobserved confounders and make use of available information. Figure extracted from [30].

The authors proceed to experimentally verify that this data-fusion approach, applied to Thompson Sampling, results in significantly less regret than competitive MABUC algorithms. The (conventional) gold standard for dealing with unobserved confounders involves randomised control trials (RCTs) [32], especially useful in medical drug testing, for example. As we noted in the data-fusion and earlier MABUC processes, randomisation of treatments may yield population-level treatment outcomes but can fail to account for individual-level characteristics. The authors provide a motivating example in the domain of *personalised medicine*, or the *effect of the treatment on the treated*, motivated by Stead et al.'s [33] observation that "despite their attention to evidence, studies repeatedly show marked variability in what healthcare providers actually do in a given situation." They proceed to formalise the existence of different treatment policies of actors in confounded decision-making scenarios. This new theory is applied to generalise RCT procedures to allow recovery of individualised treatment effects from existing RCT results. Further, they present an online algorithm which can recommend actions in the face of multiple treatment opinions in the context of unobserved confounders. For the sake of clarity, the motivating example is now briefly presented. The reader is encouraged to refer to the source material [26] for further details.

Consider two drugs which appear to be equally effective under an FDA-approved RCT. In practice, however, one physician finds agreement with the RCT results while another does not. Let us consider two potential unobserved confounders - socioeconomic status (SES) and the patient's treatment request (R). Crucially, we note that juxtaposing observational and experimental data fails to reveal these *invisible* confounders. Key to this confounded decision making (CDM) scenario is that the deciding agent (physician) do not possess a fully specified causal model (in the form of an SCM). This formalism was used to define a regret decision criterion (RDC) for optimising actions in the face of unobserved confounders. In the physician motivating example we just discussed, the intent-specific recovery rates of the first physician do not appear to differ from the observational or interventional recovery rates. The results of RDC for the second physician is only slightly off the data, at 72.5%. What is happening here? The key here is the heterogeneous intents of the agents. We now develop theory to account and exploit information implicitly contained in the multiple intents of agents.

Definition 4.9 (Heterogeneous Intents [26]). *Let A_1 and A_2 be two actors within a CDM instance, and M^{A_1} be the SCM associated with the choice of policies of A_1 and likewise M^{A_2} of A_2. For any decision variable $X \in \Pi_M$ and its associated intent $I = f_x$, the actors are said to have heterogeneous intent in $f_I^{A_1} \in F_M^{A_1}$ and $f_I^{A_2} \in F_M^{A_2}$ are distinct.*

Acknowledging the possibility of heterogeneous intents in deciding actors (such as second opinions), we can extend the notion of SCMs. Figure 12 shows a model for combining intents of different agents.

Definition 4.10 (HI Structural Causal Model (HI-SCM) [26]). *A heterogeneous intent structural causal model $M^{\boldsymbol{A}}$ is an SCM that combines the individual SCMs of actors $\boldsymbol{A} = \{A_1, \ldots, A_a\}$ such that each decision variable in $M^{\boldsymbol{A}}$ is a function of each actors' individual intents.*

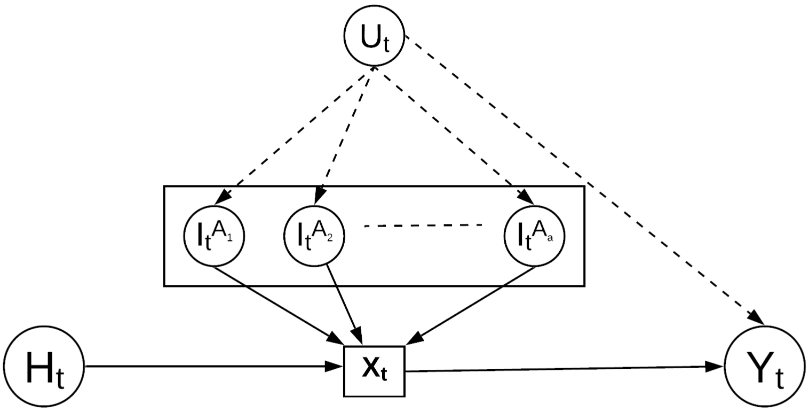


Figure 12: Figure showing example of HI-SCM. Here multiple intents of different actors contribute to the decision variable X_t. Unobserved confounder U_t affects both the intents and the outcome. The agent history is encoded as H_t. Recreated based on figure 1 in [26].

Of course, it would be naive to assume every actor adds *valuable* information to the total knowledge of the system. With this is mind, we develop the notion of an intent equivalence class.

Definition 4.11 (Intent Equivalence Class (IEC) [26]). *In a HI-SCM M^A, we say that any two actors $A_i \neq A_j$ belong to separate intent equivalence classes $\mathbf{\Phi} = \{\phi_1, \dots, \phi_m\}$ of intent functions f_I if $f_I^{A_i} \neq f_I^{A_j}$.*

With this definition in place, we can cluster equivalent actors - and their associated intents - by their IEC. For example, given $\phi_1 = \{A_1, A_2\}$ then $P(Y_x \mid I^{A_1}) = P(Y_x \mid I^{A_2}) = P(Y_x \mid I^{A_1}, I^{A_2}) = P(Y_x \mid I^{\phi_1})$. It turns out that IEC-specific optimisation is always at least equivalent to each actor's individual optimal action.

Theorem 4.7 (IEC-Specific Reward Superiority [26]). *Let X be a decision variable in a HI-SCM M^A with measured outcome Y, and let I^{ϕ_i} and I^{ϕ_j} be the heterogeneous intents of two distinct IECs ϕ_i, ϕ_j in the set of all IECs in the system, $\mathbf{\Phi}$. Then*

$$\max_{x \in X} P(Y_x \mid I^{\phi_i}) \leq \max_{x \in X} P(Y_x \mid I^{\phi_i}, I^{\phi_j}) \quad \forall \phi_i, \phi_j \in \mathbf{\Phi}.$$

Proof. WLOG, consider the case of a binary intents X. Let $x^* = \arg\max_{x \in X} P(Y_x \mid I^{\phi_i} = i^{\phi_i})$. Then

$$P(Y_x^* \mid I^{\phi_i} = i^{\phi_i}) > P(Y_x' \mid I^{\phi_i} = i^{\phi_i})$$

$$\begin{aligned} \implies & \sum_{i^{\phi_j}} P\left(Y_{x^*} \mid I^{\phi_i} = i^{\phi_i}, I^{\phi_j} = i^{\phi_j}\right) P\left(I^{\phi_j} = i^{\phi_j} \mid I^{\phi_i} = i^{\phi_i}\right) \\ > & \sum_{i^{\phi_j}} P\left(Y_{x'} \mid I^{\phi_i} = i^{\phi_i}, I^{\phi_j} = i^{\phi_j}\right) P\left(I^{\phi_j} = i^{\phi_j} \mid I^{\phi_i} = i^{\phi_i}\right) \end{aligned}$$

Letting $p = P\left(I^{\phi_j} = i^{\phi_j} \mid I^{\phi_i} = i^{\phi_i}\right)$, we can rewrite the above inequality as

$$a(p) + b(1-p) > c(p) + d(1-p)$$

(corrected mistake from proof in appendix of source). We can write it as such since we are considering the binary case. Thus if we have one case occurs, necessarily the other doesn't. We can then exhaust the cases:

1. $p = 0 \implies b > d.$
2. $p = 1 \implies a > c.$
3. $p \in (0,1) \implies \max(a,b) \geq a(p) + b(1-p).$

Thus, in each case we have that the HI-specific rewards are greater than or equal to the homogeneous-intent-specific rewards. □

With this important result in place, we can develop new criteria for decision making in a CDM with heterogeneous intents.

Definition 4.12 (HI Regret Decision Criteria (HI-RDC) [26]). *In a CDM scenario modelled by a HI-SCM M^A with treatment X, outcome Y, actor intended treatments I^{A_i}, and a set of actor IECs $\mathbf{\Phi} = \{\phi_1, \dots, \phi_m\}$, the optimal treatment $x^* \in X$ maximises the IEC-specific treatment outcome. Formally: $x^* = \arg\max_{x \in X} P(Y_x \mid I^{\phi_1}, \dots, I^{\phi_m})$.*

The authors point out that HI-RDC requires knowledge of IECs, which are not always obvious. This motivates the need for empirical means of clustering the actors into equivalence classes. Since these intents are observational (they are indicated by what naturally occurs), sampling and grouping by the following criteria suffices.

Theorem 4.8 (Empirical IEC Clustering Criteria [26])**.** *Let A_i, A_j be two agents modelled by a HI-SCM, and let their associated intents be I^{A_i}, I^{A_j} for some decision. We cluster agents into the same IEC, $\{A_i, A_j\} \in \phi_r$, whenever their intended actions over the same units correlate. In other words, if their intent-specific treatment outcomes will agree. Correlation indicated by ρ, as is common in statistics literature. Formally:*

$$\begin{aligned} \rho(I^{A_i}, I^{A_j}) = 1 &\implies \{A_i, A_j\} \in \phi_r \in \mathbf{\Phi} \\ &\implies P(Y_x \mid I^{A_i}) = P(Y_x \mid I^{A_i}, I^{A_j}) \end{aligned}$$

The authors argue that this condition can be too strict in practice and should be softened to allow for some actor-specific noise. Applying HI-RDC and empirical IEC clustering directly to the online recommendation system presented in [21] is possible but not necessarily always practical or recommended because (1) the ethics of exploring different treatment options is not always clear and (2) if UCs are present in the system and the treatment has passed experimental testing, this implies that the UCs have gone unnoticed and we - the data analysts - wouldn't necessarily know to look for them there. This motivates the need for an extension to RCTs to involve heterogeneous intents.

Definition 4.13 (HI Randomised Control Trial (HI-RDC) [26])**.** *Let X be the treatment of the RCT in which all participants of the trial are randomly assigned to some experimental condition. In other words, they have been intervened on via $do(X = x)$ with some measured outcome Y. Let $\mathbf{\Phi}$ be the set of all IECs for agents in the HI-SCM M^A for which the RCT is meant to apply. A Heterogeneous Intent RCT (HI-RCT) is an RCT wherein treatments are randomly assigned to each participant but, in addition, the intended treatments of the sampled agents are collected for each participant.*

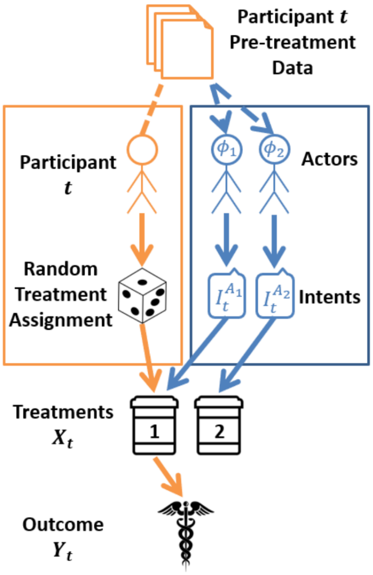


Figure 13: Figure showing the HI-RDC procedure. The traditional RCT procedure is found by following the orange path. HI-RDC adds an additional layer of actor-intent collection over and above the traditional RCT procedure. Figure extracted form [26].

This procedure yields actor IECs as well as experimental, observational, counterfactual and HI-specific treatment effects. Finally, we can implement a criterion which reveals confounding beyond what a simple mix of observational and interventional data can expose.

Theorem 4.9 (HI-RDC Confounding Criteria [26])**.** *Consider a CDM scenario modelled by a HI-SCM M^A with treatment X, outcome Y, actor intended treatments I^{A_i}, and a set of actor IECs $\Phi = \{\phi_i, \ldots, \phi_m\}$. Whenever there exists some $x \in X, i^\Phi \in I^\Phi : P(Y_x) \neq P(Y_x \mid i^\Phi)$ in M^A, then there exists some unobserved U such that $X \leftarrow U \rightarrow Y$.*

In addition to the above theory, [26] provides a procedure for an agent to attempt to repair harmful influences of unobserved confounders in an online procedure. Once again we come across the Multi-Armed Bandit with Unobserved Confounders (MABUC) in the attempt to maximise the reward (recovery in the physician example) - or minimise the regret - of the decision making process. One problem that arises in the online case is that the IECs the agent learns are not necessarily exhausitive. Let us consider whether we can find a mapping $\Phi_{on} \rightarrow \Phi_{off}$, representing a map from the online to the offline IEC sets respectively. If we can, the HI-RDC reveals optimal treatment. If not, HI-RCT data can be used to accelerate learning by gathering a *calibration unit set* - a small sample questionnaire in which actors are asked to provide intended treatments. In the case where HI-RCT IECs correspond to the sampled offline IECs, a mapping can be made. This acts as a sort of 'bootstrap' to the HI-RCT procedure by serving as a procedure to collect initial conditions for the learning process.

Definition 4.14 (Actor Calibration-Set Heuristic [26])**.** *A collection of some $n > 0$ calibration units from an offline HI-RDC dataset $\mathcal{D}$ can be used to learn the IECs of agents in an online domain. Three heuristics scores $h(t) = h_c(t) + h_d(t) + h_o(t)$ can be used to guide the selection procedure:*

1. ***Consistency:*** *how consistent agents in the same IEC $\phi_r = \{A_1, \dots,_i\}$ are with their intended treatment, $h_c(t) = (Number\ of\ I_t^A \in \phi_r\ agreeing\ with\ majority)/|\phi_r|$.*
2. ***Diversity:*** *how often a configuration of I^Φ has been chosen, favouring a diverse set of IEC intent combinations, $h_d(t) = 1/(Number\ of\ times\ I^\Phi\ appears\ in\ calibration\ set)$. This acts by favouring 'exploration' in terms of the IEC space.*
3. ***Optimism:*** *a bias towards choosing units in which the randomly assigned treatment x_t was optimal and succeeded or suboptimal and failed,*

$$\begin{aligned} h_o(t) = &\left(P\left(Y_{x_t} \mid I_t^\Phi\right) > P\left(Y_{x'} \mid I_t^\Phi\right) \forall x' \in X \backslash x_t\right) 1\left(Y_t = 1\right) \\ &+ \left(P\left(Y_{x_t} \mid I_t^\Phi\right) < P\left(Y_{x'} \mid I_t^\Phi\right) \exists x' \in X \backslash x_t\right) 1\left(Y_t = 0\right) \end{aligned}$$

The calibration set is given by $h(\mathcal{D}, n) = \{t \in \mathcal{D} \mid n\ highest\ h(t)\}$.

The authors experimentally show that agents maximising HI-specific rewards, clustering by IEC, and using calibration sets with and without the Actor Calibration-Set Heuristic, each outperforms the previous version. In other words, each step improves upon the regret the agent experiences. When compared against an *oracle* - that treated UCs as observed (unrealisable) - the agent performs well. [34] asks some interesting questions and presents intriguing results. They tackle the problem of an agent and humans having different sensory capabilities, and thereby "disagreeing" on their observations. The authors find that when leaving human intuition out of the loop - even when the agent's sensory abilities are superior - results in worse performance. The theory presented in this section is sufficient to understand the presentation of the results in [34], and the reader is encouraged to engage with the source.

Now that we have considered counterfactual decision making within a causal system, we should consider how we can transfer and generalise causal results across domains. This is what the next section tackles.

4.4 Task 4: Generalisability and Robustness

One of the most important features of human intelligence is its ability to generalise and transfer causal knowledge across seemingly disparate domains. This allows powerful inferences and decision making procedures possible even in foreign environments. Bareinboim and Pearl address the problem of transferring knowledge from data collected in heterogeneous domains $\Pi = \{\pi_1, \dots, \pi_n\}$ to some target domain π^* - a problem known as mz-transportability. In [35] the authors establish a necessary and sufficient condition for deciding whether this transfer is feasible. In the sciences, powerful studies involving transfer of knowledge across related domains are known as *meta-analysis* or *externally valid* studies. The transfer of causal knowledge is known as *transportability* and is a crucial ability needed for artificial agents to automate the process of knowledge acquisition, discovery and learning.

Consider an example in which we would like to use knowledge of social science experiments done in Los Angeles (predicting outcome Y with cause X, confounded by some age distribution Z) to make similar predictions in New York. Calling the distribution in Los Angeles $P(y \mid do(x))$, we would like to predict $R = P^*(y \mid do(x))$ - the cause/effect relationship under a different age distribution in New York. We call the process which generates this difference in age across the populations a *difference generating factor*, denoted graphically by ■, which are caused by some set of *selection variables* S. In this case we have $S \rightarrow Z$. We can then derive an remarkably simple *transport formula* (3) as follows:

$$\begin{aligned} R &= \sum_s P^*(y \mid do(x), z) P^*(z) \\ &= \sum_s P(y \mid do(x), z) P^*(z) \end{aligned} \tag{3}$$

This deceptively simple formula tells us we can estimate R - an interventional quantity - using a *drop in* observational distribution $P^*(z \mid do(x), z)$. This acts to re-weight observations by the interventional affect in a different domain. To generalise a transport formula is not completely trivial. We need to know whether *do*-calculus is complete. That is, whether *do*-calculus operations can always find such a transport formula. Recall that causal models and induced diagrams encode relationships under a particular domain. A formalism that is helpful for the study of transfer of knowledge across causal domains is the notion of selection diagrams. These diagrams graphically encode the shared causal relations and difference generating factors of different causal systems.

Definition 4.15 (Selection Diagrams [35]). *Let $\langle M, M^* \rangle$ be a pair of structural causal models from domains $\langle \pi, \pi^* \rangle$, sharing causal diagram G. The pair $\langle M, M^* \rangle$ induces a selection diagram D if D obeys the following criteria: (1) every edge in G is also an edge in D, and (2) D contains an extra edge $S_i \to V_i$ whenever there might exist some discrepancy $f_i \neq f_i^*$ or $P(U_i) \neq P^*(U_i)$ between M and M^*.*

These S variables in the selection diagram serve as identifying the mechanisms where structural differences in the data generating process takes place between models under different domains. Knowledge of these structural overlaps of different causal domains allows us to formalise what it means to transfer knowledge between domains. This is the notion of mz-transportability discussed earlier. Simply put, knowledge is transferable between domains only if the causal effect R can be determined from information available in the observational and interventional distributions.

Definition 4.16 (mz-Transportability [35]). *Let $\mathcal{D} = \{D^{(1)}, \ldots, D^{(n)}\}$ be a collection of selection diagrams with source domains $\Pi = \{\pi_1, \ldots, \pi_n\}$ and target domain π^*. Let $\boldsymbol{Z}_i$ be the variables in which experiments can be conducted in domain π_i. If $\langle P^i, I_z^i \rangle$ are the observational and interventional distributions, then the causal effect $R = P_{\boldsymbol{x}}^*(\boldsymbol{y})$ is said to be mz-transportable from Π to π^* in $\mathcal{D}$ if $P_{\boldsymbol{x}}^*(\boldsymbol{y})$ is uniquely computable from $\cup_{i=1,\ldots,n} \langle P^i, I_z^i \rangle \cup \langle P^*, I_z^* \rangle$ in any model that induces $\mathcal{D}$.*

The above graphical condition has a counterpart that can be written in terms of *do*-calculus criteria.

Theorem 4.10. *Let symbols be defined as above. The effect $R = P^*(\boldsymbol{y} \mid do(x))$ is mz-transportable from Π to π^* if the expression $P(\boldsymbol{y} \mid do(x), \boldsymbol{S}_1, \ldots, \boldsymbol{S}_n)$ is reducible using the rules of do-calculus to an expression in which (1) do-operators that apply to subsets o I_z^i have no $\boldsymbol{S}_i$-variables or (2) do-operators apply only to subsets of I_z^i.*

This theorem tells us that do-calculus is complete in terms of finding these transport formulae. The authors also prove completeness for an established algorithm for computing transport formulae. Refer to source material [35] for details of this algorithm.

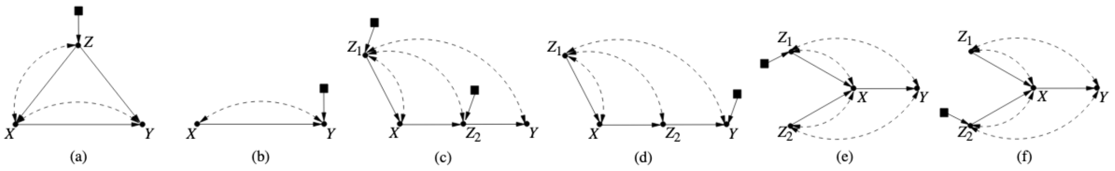


Figure 14: Figures (a) through (f) show illustrative examples of transportability in causal selection diagrams. These highlight the important of the nature of unobserved confounders. (a) This diagram shows an example of when transportability of $R = P^*(y \mid do(x))$ is trivially solved by re-weighting of the variable directly affected by difference-generating variable. In this case $S \to Z$. (b) shows the simplest example in which one cannot transport a causal relation between domains. Even by randomisation on X, the causal effect is not uniquely computable due to UCs. (c) and (d) show examples where transportability of causal effects require interventional information over Z_1 in π_1 and Z_2 in π_2, but not over $\{Z_1, Z_2\}$ in the combined domain. (e) and (f) show examples where transportability is only possible in the combined domain. Figure extracted from [35].

This process of transferring knowledge relates well to the concept of unifying big data. The ability to fuse multiple datasets, collected under heterogeneous conditions, without incurring large bias penalties is something critically important for generalising an agent's ability to learn under different conditions. In [11] Bareinboim and Pearl review this problem of data fusion under the auspice of causal inference. In [36] Lee et al. argue that *identifiability* and *randomisation* are two extremes in approach to inferring cause-effect relationships from some combination of observations, experiments and prior (substantive) knowledge. In fact, *z-identifiability* (zID) generalises exactly this question for the case where all possible interventions (experiments) are possible. The authors argue that this requirement is (obviously) not always reasonable and propose a generalisation such that any expression derivable from an arbitrary collection of observations and experiments is returned by the proposed algorithm. The following theory is developed to introduce a strategy that is used to prove non-gID (defined later) which allows for a graphical, necessary and sufficient condition for the causal decision problem of interest. We start by defining a c-component.

Definition 4.17 (C-component [37]). *Let causal graph $\mathcal{G}$ be such that a subset of its bidirected arcs forms a spanning-tree over all its vertices, then $\mathcal{G}$ is a confounded-component (c-component).*

With this definition in mind, notice the c-components in figure 14. We use $\mathcal{C}(\mathcal{G})$ to denote the set of c-components that partitions the vertices in $\mathcal{G}$ such that $\mathcal{C}(\mathcal{G}) = \{\boldsymbol{W}_i\}_{i=1}^{k}$ implies that $\mathcal{G}[\boldsymbol{W}_i]$ is a c-component for every $\boldsymbol{W}_i \subseteq \boldsymbol{V}$, the endogenous (visible) variables.

Definition 4.18 (C-forest [36]). *A causal graph $\mathcal{G}$ with root set $\boldsymbol{R}$ is an $\boldsymbol{R}$-rooted c-forest if $\mathcal{G}$ is a c-component with minimal number of edges.*

All of figure 15(a) through 15(b) are c-components since there are unobserved confounders (bidirected edges) spanning the vertices. Further 15(a) through 15(c) are c-forests since they have minimal number of spanning bidirected edges.

Definition 4.19 (Hedge [36]). *A hedge is a pair of $\boldsymbol{R}$-rooted c-forests $\langle \mathcal{F}, \mathcal{F}' \rangle$ such that $\mathcal{F}' \subseteq \mathcal{F}$.*

By this definition, figure 15(a) and 15(b) are hedges because we can find two c-forests $\mathcal{F}$ and $\mathcal{F}'$ such that $\mathcal{F}' \subseteq \mathcal{F}$. Crucially, 15(c) is not a hedge since the spanning bidirected edges are not minimal. This type of structure prevents g-identifiability, which is now formalised and discussed.

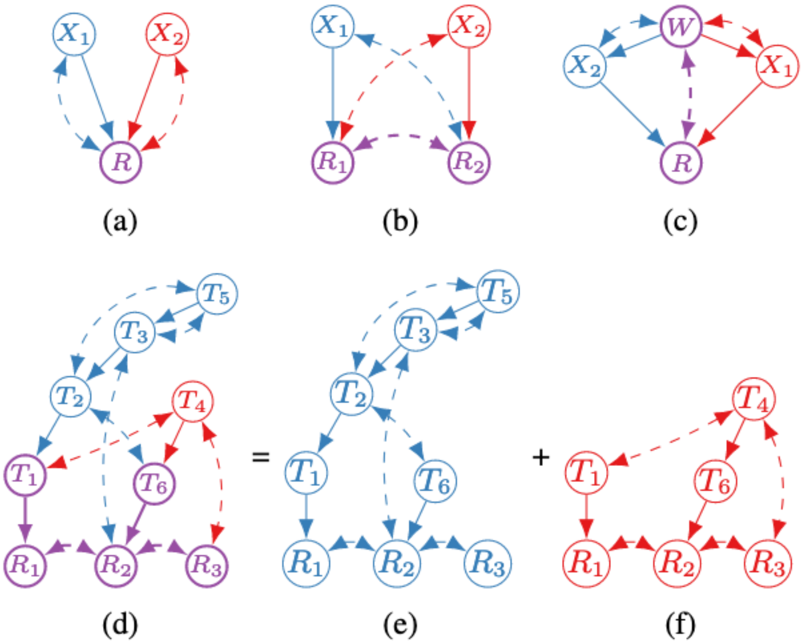


Figure 15: Figure showing examples of hedges, c-components, c-forests, and thickets. These form graphical criteria for g-identifiability. Details are discussed in the text itself. Thickets are shown to preclude g-identifiability. Crucially, (d) is shown to be an overlap of hedges which forms a thicket. Figure extracted from [36].

Definition 4.20 (g-Identifiability [36]). *Let $\boldsymbol{X}, \boldsymbol{Y}$ be disjoint sets of variables, $\mathbb{Z} = \{\boldsymbol{Z}_i\}_{i=1}^{m}$ be a collection of sets of variables, and let $\mathcal{G}$ be a causal diagram. If $P_x(y)$ is uniquely computable from distributions $\{P(\boldsymbol{V} \mid do(z))\}_{\boldsymbol{Z}\in\mathbb{Z}, \boldsymbol{z}\in dom(\boldsymbol{Z})}$ in any causal model which induces $\mathcal{G}$, we say that $P_x(y)$ is g-identifiable from $\mathbb{Z}$ in $\mathcal{G}$. Here $P(\boldsymbol{V})$ is the probability distribution describing the natural state of the system (assumed to be available).*

Simply put, we say the distribution is g-identifiable with respect to a set of intervenable variables in the causal system if they are sufficient to uniquely compute it. This set of variables are the ones we intervene on by doing an experiment, as we discussed earlier. In this way it is a generalisation of z-identifiability discussed earlier. We now introduce some more definitions needed for the non-gID criteria.

Definition 4.21 (Hedgelet decomposition [36]). *The hedgelet decomposition of a hedge $\langle \mathcal{F}, \mathcal{F}' \rangle$ is the collection of hedgelets $\{\mathcal{F}(\boldsymbol{W})\}_{\boldsymbol{W}\in\mathcal{C}(\mathcal{F}'')}$ ($\mathcal{F}'' = \mathcal{F} \setminus \mathcal{F}'$) where each hedgelet $\mathcal{F}(\boldsymbol{W})$ is a subgraph of $\mathcal{F}$ made of (i) $\mathcal{F}[\boldsymbol{W}(\mathcal{F}) \cup \boldsymbol{W}]$ and (ii) $\mathcal{F}[De(\boldsymbol{W}_{\mathcal{F}})]$ without bidirected edges.*

Referring back to figure 15, some possible hedgelet decompositions are colour coded in blue and red to indicate distinct hedgelets, with purple used to indicate the shared variables (commonly root sets). This leads us nicely to the last definition we need for this criterion. Though this definition appears arbitrarily technical, it is rather intuitive once the reasoning is developed.

Definition 4.22 (Thicket [36]). *Let $\boldsymbol{R}$ be non-empty set of variables and $\mathbb{Z}$ be a collection of sets of variables in $\mathcal{G}$. A thicket $\mathcal{J} \subseteq \mathcal{G}$ is an $\boldsymbol{R}$-rooted c-component consisting of a minimal c-component over $\boldsymbol{R}$ and hedges*

$$\mathbb{F}_{\mathcal{J}} = \{\langle \mathcal{F}_{\boldsymbol{Z}}, \mathcal{J}[\boldsymbol{R}] \rangle \mid \mathcal{F}_{\boldsymbol{Z}} \subseteq \mathcal{G} \setminus \boldsymbol{Z}, \boldsymbol{Z} \cap \boldsymbol{R} = \varnothing\}_{\boldsymbol{Z}\in\mathbb{Z}}.$$

Let's consider this definition step-by-step by considering figure 15(c). First, we notice the graph is a c-component that *contains* a minimal c-component. It does not necessarily need to be a c-forest itself. Next, we need a pair of $\boldsymbol{R}$-rooted c-forests $\langle\mathcal{F}_{\boldsymbol{Z}}, \mathcal{J}[\boldsymbol{R}]\rangle$. We select the graphs induced by sets $\{W, X_1, R\}$ and $\{W, X_2, R\}$ with $\boldsymbol{Z} = \{\{X_1\}, \{X_2\}\}$ the intervention set. Then we have hedges $\langle\mathcal{F}_{X_1}, \mathcal{J}[R]\rangle$ and $\langle\mathcal{F}_{X_2}, \mathcal{J}[R]\rangle$ that overlap and have intervention variables $Z \in \boldsymbol{Z}$ that do not intersect with the root set, $\boldsymbol{R} = \{R\}$. Basically, a thicket is an overlapping of hedges, and hedges were the 'bad' structure that prevented gID in the causal graph. Though this is a fairly involved procedure to do manually, especially on large causal graphs, it is algorithmically feasible as shown in Lee et al. The usefulness of this algorithm relies on the following result.

Theorem 4.11 (Thicket non-gID [36])**.** *If there exists some thicket $\mathcal{J}$ for $P_{\boldsymbol{x}}(\boldsymbol{y})$ in causal graph G with respect to intervention set $\mathbb{Z}$, then $P_{\boldsymbol{x}}(\boldsymbol{y})$ is not g-identifiable in G.*

To make this idea explicit we include the following figure 16 extracted from slides provided directly by Sanghack Lee, coauthor of several papers (including [36]) presented in this work [38].

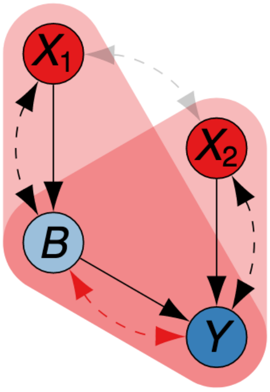


Figure 16: Thicket structure for $P_x(y)$ identified as an overlap of distinct hedges, each colours as a red rounded triangle. Extracted from [38].

This completes the required formalism's for identifying structural constraints from explicit causal models. This ties well into the next section in which we discuss how we can apply this theory to learn causal structure from observational and interventional data. This is especially for allowing reinforcement learning agents to uncover causal structure.

4.5 Task 5: Learning Causal Models

Perhaps one of the most computationally difficult processes in the field of causal inference is that of learning underlying causal structure by algorithmically identifying cause-effect relationships. In recent years there has been a surge of interest in learning such relationships in the fields of machine learning and artificial intelligence, though it has been relatively prevalent in the social sciences for many years now (e.g. [39] has over 25000 citations at time of writing). The discovery of IC and PC [40] algorithms independently displayed the feasibility of learning such causal structure from observational data - a fact that was not obvious at the time. Since these discoveries new methods of inferring such structure have emerged. Many of these methods require satisfaction of the strict *causal sufficiency* assumption. This requires that no latent variables affects more than one observed variable. In other words, these algorithms do not deal with confounding. [41] improves upon previous work by introducing an algorithm that can learn any causal graph, as well as the existence and location of the latent variables using $\mathcal{O}(d\log(n) + l)$ interventions, where d is the largest node degree, and l is the longest directed path of the causal graph. Further, they introduce a probabilistic algorithm which can learn the observable graph and all the latent variables using $\mathcal{O}(d\log^2(n) + d^2\log(n))$ interventions with high probability. We discuss and develop this theory as it is deemed a particularly interesting approach to the problem at hand. The authors split the task of learning the observable graph and latent variables into three distinct sub-tasks. They start by proposing a method for finding the *transitive closure* of the observable graph. Next, this transitive closure is *reduced* to reveal some subset of the edges in the underlying causal graph. Conditional independence tests are then used to uncover latent variables. We now discuss select theory in detail.

The authors begin by showing that *separating systems* can be used to construct sequences of pairwise conditional independence tests to discover the transitive closure of the observable causal graph. That is, to discover the causal paths in the causal system by testing which variables 'rely' on other variables (in an informal sense). To develop this idea formally we require the idea of a post-interventional causal graph. This is simply the causal graph G with all edges directed onto intervened variables, removed. Recall that faithfulness indicates that causal relations are only formed as a result of d-separation. Simply put, there are no relationships that perfectly balance each other so as to appear to have no causal relations. These conditions allow us to formalise the conditional independence test we require.

Lemma 4.1 (Pairwise Conditional Independence Test [41])**.** *Consider causal graph with latent variables D_l, and an intervention set $S \subset V$ of observable variables. Applying the post-interventional faithfulness assumption, we have that for any pair $X_i \in S, X_j \in V \setminus S, (X_i \not\perp\!\!\!\perp X_j)_{D_l[S]}$ if and only if X_i is an ancestor of X_j in the post-interventional graph $D[S]$.*

This lemma provides a method for determining ancestry for any ordered pair of variables, (X_i, X_j). Crucially though, this method is not sufficient. For example, consider $X_i \to X_k \to X_j$ where $X_i, X_k \in S, X_j \notin S$. The authors propose resolving this issue by using a sequence of interventions guided by a separating system. The correct causal graph can then be learned by finding the transitive closure.

Definition 4.23 ((m,n) Strongly Separating System [41])**.** *An (m,n) strongly separating system is a collection of subsets $\{S_1, S_2, \dots, S_m\}$ of the ground set $[n]$ such that for any two pairs of nodes i and j, there exists a set S in the family such that $i \in S, j \notin S$ and also another set S' such that $i \notin S', j \in S'$.*

This definition is useful because, as shown in [41], a strong separating system always exists on ground set $[n]$ when $m \leq 2\lceil \log n \rceil$. This allows us to introduce a deterministic algorithm for learning the observable causal graph D from the ancestral relationships, which requires only $2\lceil \log n \rceil$ interventions and conditional independence tests. A key insight for the deterministic algorithm is that whenever the intervention set contains all parents of X_i, the only variables that are dependent with X_i in the post-interventional set are the parents, Pa_i. Consider r, the longest directed path of D_{tc}. Using the obvious partial order $<_{D_{tc}}$ on the vertex set V, we can define a unique partitioning of vertices $\{T_i \mid i \in [r+1]\}$ where $T_i <_{tc} T_j \forall i < j$. Each node in i is thus a set of of mutually incomparable elements and represents the set of nodes at layer i in the transitive closure graph D_{tc}. Define $\mathcal{T}_i = \cup_{k=1}^{i-1} T_k$, then $Pa_i \subset T_i$ - a fact that is exploited for the deterministic algorithm the authors present. Perhaps the most interesting aspect of this paper is the randomised algorithm the authors propose. The strategy employed here is to repeatedly use the ancestor graph learning algorithm to learn the observable graph. This procedure makes use of transitive reduction.

Definition 4.24 (Transitive Reduction)**.** *Given a directed acyclic graph $D = (V, E)$, let its transitive closure be D_{tc}. Then $Tr(D) = (V, E_r)$ is a directed acyclic graph with minimum number of edges such that its transitive closure is identical to D_{tc}.*

This transitive reduction is simple and effective. This allows for an iterative procedure for revealing causal relationships. We now elaborate on this procedure in the following lemma, discussed in [41].

Lemma 4.2. *Given intervention set $S \subset V$ of nodes in the observable causal graph D, we can notate the post-interventional observable causal graph as $D[S]$. Now, consider a specific observable node $V_i \in S^C$, and let Y be a direct parent of V_i in D such that all the direct parents of V_i above Y in the partial order $\pi()$ are in S. Formally, $\{X \mid \pi(X) > \pi(Y), (X, V) \in D\} \subseteq S$. Then $Tr(D[S])$ will contain the directed edge (Y, V_i) and it can be computed from $Tr((D[S])_{tc})$.*

To solidify the simplicity of the lemma, consider figure 17. Here we show how an intervention changes what the procedure can reveal about the structure of the transitive relationships in the observable causal graph. Refer to the image caption for details about what each successive graph represents.

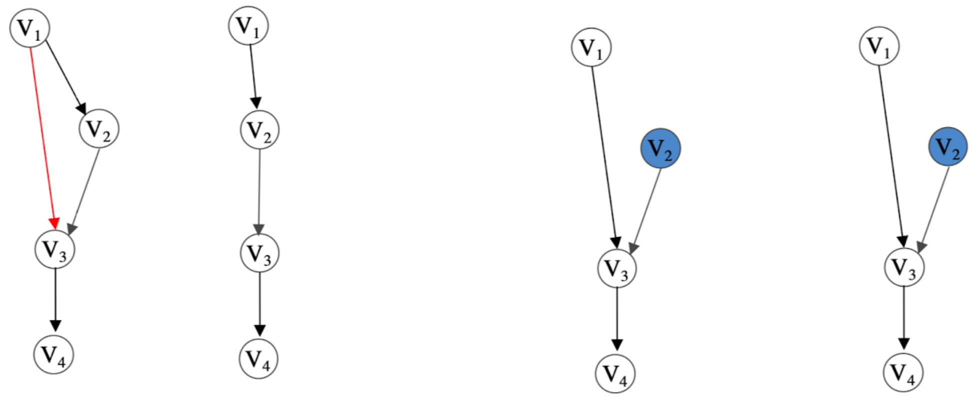


Figure 17: Figure showing examples of theory in [41]. Starting from the left, we have an example of a graph without latent variables. Next, we have the result of the transitive reduction procedure on the previous graph. Notice that the red edge has not been revealed. Next, we have the same observational graph but with V_2 intervened on. This removes the direct causal relation between V_1 and V_2 and thus reveals edge $V_1 \to V_3$ in the transitive reduction procedure (shown last). Figure extracted from [41].

This procedure is useful for the probabilistic algorithm the authors propose. Here, the basic idea is that random intervention and transitive closure computation can reveal edges of the underlying causal graph. As we perform more interventions, our certainty about the structure of the observable causal graph rises. The following theorem [41] formalises this idea.

Theorem 4.12. *Let d_{max} exceed the the maximum in-degree in the observable graph D. The proposed probabilistic algorithm requires at most $8cd_{max}(\log n)^2$ interventions and conditional independence tests on samples obtained from post-interventional distributions, and returns the observable portion of the causal graph with a minimum probability of $1 - \frac{1}{n^{c-2}}$.*

We now consider a more involved scenario. Recall that conditioning on an observable is not necessarily sufficient due to backdoor paths. Consider the scenarios in 18. On the left, we have an example of a graph where intervening on X leaves an influencing path open, $X \leftarrow U \rightarrow T \leftarrow M \rightarrow Y$. This means that observation does not necessarily match our intervention data, $P(Y \mid X) \neq P(Y \mid do(X))$. We must intervene on a parent of X to remove confounding influences in this example. Similarly for the graph on the right where we need to intervene on parents of Y. This motivates the following theorem.

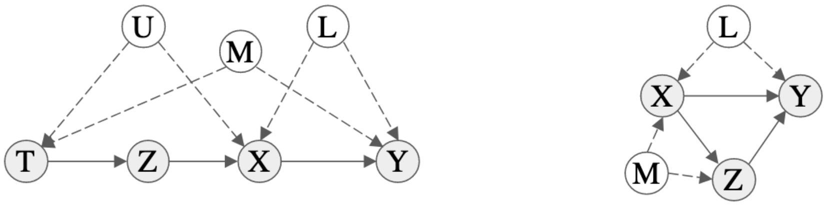


Figure 18: Figure showing examples of graphs that require a more complex intervention to block backdoor paths. Details about these are discussed in the text. Figure extracted from [41].

Theorem 4.13 (Interventional Do-See Test [41])**.** *Given a causal graph G over observable variables V and latents L. Denoting edge set of G by E, then $P(V_j \mid V_i = v_i, do(Pa_i = pa_i, Pa_j = pa_j)) = P(V_j \mid do(V_j \mid do(V_i = v_i, Pa_i = pa_i, Pa_j = pa_j)))$, if and only if there exists some $k \in \mathbb{N}$ such that $(L_k, V_i) \in E$ and $(L_k, V_j) \in E$. Recall, Pa_i are the parents of V_i.*

The final result of this paper is, perhaps, the most mathematically satisfying and, otherwise, surprising. We will need some theory of graph colourings to present this.

Definition 4.25 (Strong Edge Colouring [41])**.** *A strong edge colouring of a (undirected) graph with k colours is a mapping of edges to a colour class, $\chi : E \rightarrow [k]$, such that any two distinct edges that are incident on adjacent vertices have different colours assigned to them.*

This definition leads to the following result [42].

Lemma 4.3. *Given a graph G with maximum degree d, we can strongly edge colour G using at most $2d^2$ colours. Simply apply a greedy algorithm to colouring edges in sequence.*

Remarkably, only two interventions are required per colour class for the do-see interventional test 4.13. That is, one intervention for the 'do' part and one for the 'see' part. The authors exploit this and the following theorem to present an efficient algorithm for learning the latent edges of an observable graph with maximum degree d.

Theorem 4.14. *At most $4d^2$ interventions are required to learn the latent variables in the observable graph.*

[43] extends the work in causal structure learning by focusing on limiting the interventions to non-adaptive experiments of unit size. The authors show a greedy algorithm achieves a $(1 - \frac{1}{e})$-approximation to the optimal objective. It is well understood that whenever it is possible to perform a sufficient number of interventions, the underlying causal structure of a system can be fully recovered. The authors propose focusing on the question: "for a fixed budget of interventions, what portion of the causal graph is learnable?" This question is of interest because, in some applications, performing simultaneous interventions on multiple variables is not possible. This relates well to our discussion of z and g-identifiability. Recall two DAGs are Markov equivalent if they share conditional independence results. This allows the definition of the essential graph, which will prove useful in later results.

Definition 4.26 (Essential Graph [43])**.** *The essential graph of G, denoted $Ess(G)$, is a mixed graph (contains both directed and undirected edges) where the directed edges are the edges that have common direction for all elements in the Markov equivalence class of G. Similarly, the undirected edges are those that differ in direction for at least two elements of the equivalence class.*

It is important to clarify what we mean by experiment or intervention. For the most part these are equivalent. In this case, however, we differentiate by noting that an experiment can consist of multiple interventions. Here we wish to deal with one intervention at a time. An *interventional structure learning* algorithm consists of a set of k experiments $\mathcal{E} = \{\mathcal{E}_1, \dots, \mathcal{E}_k\}$ where each of the k experiments contains m_i interventions. The authors consider the situation $m_i = 1 \; \forall i \in \{1, \dots, k\}$. The experiment set $\mathcal{I}$ thus leads to the discovery of the orientation of the edges intersecting members in the set, denoted $A_{G^*}^{(\mathcal{I})}$ where G^* represents the true causal DAG. Letting $H = (V(H), E(H))$ denote the undirected subgraph of $Ess(G^*)$ (i.e. The subgraph with edges having disagreeing directions in the equivalence class). Further, letting $R(\mathcal{A}, G^*)$ denote the the subset of $E(H)$ that can be learned by applying Meek rules [44], then $D(\mathcal{I}, G^*) = |R(A_{G^*}^{(\mathcal{I})}, G^*)|$ represents the cardinality of the set of edges that can be learned by experiment set $\mathcal{I}$. Finally, let

$$\mathcal{D}(\mathcal{I}) = \mathbb{E}_{G_i}[D(\mathcal{I}, G_i)] = \frac{1}{|\mathcal{G}|} \sum_{G_i \in \mathcal{G}} D(\mathcal{I}, G_i).$$

Then the problem we are interested in can be formulated as computing

$$\max_{\mathcal{I} \subseteq \mathcal{V}} D(\mathcal{I}) \quad \text{s.t. } |\mathcal{I}| = k.$$

All this formalism says is that we would like maximise the number of edges we can learn about by performing experiments of size one. The problem, however, is that finding such an intervention set, $\mathcal{I}$, requires combinatorial search, and computing $D(\mathcal{I})$ can be intractable. Dealing with this requires some theory of set functions.

Definition 4.27 (Submodularity [43])**.** *A set function $f : 2^V \to \mathbb{R}$ is submodular if for all subsets $\mathcal{I}_1 \subseteq \mathcal{I}_2 \subseteq\subseteq V$ and all $v \in V \setminus \mathcal{I}_2$, the following condition is satisfied*

$$f(\mathcal{I}_1 \cup \{v\}) - f(\mathcal{I}_1) \geq f(\mathcal{I}_2 \cup \{v\}) - f(\mathcal{I}_2).$$

Given a submodular function with $f(\varnothing) = 0$ that is monotonically increasing, then the set of interventions found by the greedy algorithm (presented below) satisfies $f(\hat{\mathcal{I}}) \geq (1 - \frac{1}{e}) \max_{|\mathcal{I}|=k} f(\mathcal{I})$ [45]. Thus, all we need to show is that the set function $\mathcal{D}$ defined earlier is monotonically increasing and submodular and the result follows. The proof of monotonicity follows fairly easily from the definitions, while the submodularity is more involved. See the supplementary materials in [43] for details.

Algorithm 4 General Greedy Algorithm

```
Result: Return near-optimal intervention set
Apply algorithm (CCI) to learn Markov equivalence class
 Obtain $Ess(G^*)$ from Markov equivalence class
 Initialise: $\mathcal{I}_0 = \varnothing$ for $i = 1, \dots, k$ do
    $v_i = \arg\max_{v \in V \setminus \mathcal{I}_{i-1}} \hat{\mathcal{D}}(\mathcal{I}_{i-1} \cup \{v\}) - \hat{\mathcal{D}}(\mathcal{I}_{i-1})$
     $\mathcal{I}_i = \mathcal{I}_{i-1} \cup \{v_i\}$
end
```

The greedy algorithm relies on the theory developed and the results presented in the appendices of [43]. This algorithm iteratively adds a variable with the greatest marginal gain to the intervention set, $\Delta_v(\mathcal{I}) = \mathcal{D}(\mathcal{I} \cup \{v\}) - \mathcal{D}(\mathcal{I})$, until the budget is exhausted. In other words, it *greedily* selects a possible intervention. The intractability of computing $\mathcal{D}(\mathcal{I})$ is addressed by proposing a Monte-Carlo approach. The proposed algorithm employs random sampling and generates multisets of DAGs, G', in the algorithm. The following result provides theoretical legitimacy to this method [43].

Theorem 4.15. *Imagine we are given some estimate of the number of interventions required to learn the about the edges in the underlying causal graph, with $\mathbb{E}[D(\mathcal{I}, G_i')] = \mathcal{D}(\mathcal{I})$. If we are given set $\mathcal{I}$ and $\epsilon, \delta > 0$, and if*

$$N = |G'| > \frac{|E(Ess(G^*))|(2+\epsilon)}{\epsilon^2} \ln(\frac{2}{\delta}), \; \textit{then} \quad \mathcal{D}(\mathcal{I})(1-\epsilon) < \hat{\mathcal{D}}(\mathcal{I}) < \mathcal{D}(\mathcal{I})(1+\epsilon),$$

with probability larger than $1 - \delta$.

Proof. Define $X = \frac{D(\mathcal{I}, G_i')}{|E(Ess(G^*))|}$, for $i \in \{1, \dots, N\}$. By the assumption of the theorem, $\mathbb{E}[X_i] = \frac{1}{|E(Ess(G^*))|}\mathcal{D}(\mathcal{I})$. Applying the Chernoff bound we have

$$\begin{aligned} P(\sum_{i=1}^{N} X_i - \frac{N}{|E(Ess(G^*))|}\mathcal{D}(\mathcal{I}) \geq \epsilon \frac{N}{|E(Ess(G^*))|}\mathcal{D}(\mathcal{I})) \\ \leq 2\exp\left(-\frac{N\epsilon^2}{|E(Ess(G^*)|(2+\epsilon)}\right)\mathcal{D}(\mathcal{I}) \\ \leq 2\exp\left(-\frac{N\epsilon^2}{|E(Ess(G^*)|(2+\epsilon)}\right). \end{aligned}$$

Therefore,

$$P(|\frac{1}{N}\sum_{i=1}^{N} D(\mathcal{I}, G'_i) - \mathcal{D}(S)|) \geq \epsilon\mathcal{D}(\mathcal{I}) \leq 2\exp(-\frac{N\epsilon^2}{)}|E(Ess(G^*))|(2+\epsilon)).$$

This further implies

$$P(|\hat{\mathcal{D}}(\mathcal{I}) - \mathcal{D}(\mathcal{I})| < \epsilon\mathcal{D}(\mathcal{I})) > 1 - 2\exp(-\frac{N\epsilon^2}{|E(Ess(G^*))|(2+\epsilon)}).$$

Applying these results, we can set $N > \frac{|E(Ess(G^*))|(2+\epsilon)}{\epsilon^2}\ln(\frac{2}{\delta})$ to obtain an upper bound of $1-\delta$. This completes the proof. □

The authors further propose accelerating the greedy algorithm by performing *lazy* evaluations. That is, they exploit the monotonicity of the marginal gains such that if $\Delta_{v_1}(\mathcal{I}_i) > \Delta_{v_2}(\mathcal{I}_i)$ and $\Delta_{v_1}(\mathcal{I}_{i+1}) > \Delta_{v_2}(\mathcal{I}_i)$, then $\Delta_{v_1}(\mathcal{I}_{i+1}) > \Delta_{v_2}(\mathcal{I}_{i+1})$. This improves performance in a similar manner to dynamic programming improvements over naive methods. These procedures are combined to empirically show that a significant portion of causal systems can be learned using only a relatively small number of interventions - a remarkable result for a seemingly intractable problem!

The problem of learning causal structure is still a very active area of research. Extension of work presented here is considered in [46] and [47] for example. For brevity these are not discussed further. We have developed the bulk of the theory necessary to actually implement some useful and promising agents. The next task develops some theory needed to apply some of these tools to the task of imitation learning.

4.6 Task 6: Causal Imitation Learning

We now reach the final task presented by Bareinboim in his development of the causal reinforcement learning framework [9]. This task involves the interesting challenge of learning by expert demonstration - imitation learning. In their now classic paper, Abbeel and Ng [48] explored applying inverse reinforcement learning (IRL) techniques to the imitation learning procedure. Essentially, IRL learns a reward function that emphasises the observed expert trajectories. This is in contrast to the other common method of imitation learning known as behaviour cloning where an agent seeks to mimic the policy of the expert. Both these methods have been successful in their own right, but they make the strong assumption that actions of the expert are fully observed by the imitator. [49] addresses some of these shortcomings by introducing a complete graphical criterion for determining whether imitation learning is feasible given observational data and knowledge about the underlying causal process. Further, a sufficient algorithm for identifying an imitation policy when this criterion does not hold is presented.

Definition 4.28 (Partially Observable SCM [49])**.** *A POSCM is a tuple $\langle M, \boldsymbol{O}, \boldsymbol{L}\rangle$. Here M is an SCM, $\boldsymbol{O}$ represents the observed endogenous variables, and $\boldsymbol{L}$ represents the latent (unobserved) endogenous variables. The observed and latent variables are mutually exhaustive over the endogenous variables.*

The task at hand is to determine the value of performing some intervention (action) that is part of the observed variable set, $X \in \boldsymbol{O}$. Assuming the reward is latent, we wish to identify a policy π such that the expected reward $\mathbb{E}[Y \mid do(\pi)]$ exceeds a certain minimum performance requirement, τ. We say $P(y \mid do(\pi))$ is identifiable if for a subset of the exogenous variables, $\boldsymbol{Y} \subseteq \boldsymbol{V}$, the distribution $P(y \mid do(\pi); M)$ is uniquely computable from the observation distribution and POSCM, M. In other words, if we can *identify* the outcome from the observations of the expert behaviour in an imitation learning context. In fact, when the reward $\boldsymbol{Y}$ is latent (not all $\boldsymbol{y} \in \boldsymbol{Y}$ are observed), we cannot identify $P(\boldsymbol{y} \mid do(\pi))$ (see corollary 1 in [49]). We thus need more information to learn an effective imitation policy.

By assuming that the observed actions are demonstrated by an expert (exceeds a threshold), we relax the need to worry about non-identifiability issues. Further, we say a reward distribution $P(\boldsymbol{y})$ is imitable if there exists some policy in the policy space that can identify the distribution for some POSCM. For example, consider $X \to W \to Y$ with policy $\pi(x) = P(x)$. Then the interventional distribution is $P(y \mid do(\pi)) = \sum_{x,w} P(y \mid x)P(w \mid x)\pi(x) = P(y)$. In other words, in this simple case the unobserved reward distribution $P(y)$ is imitable purely by observing realisations of action X. Importantly, as previously discussed, the reward distribution remains unidentifiable. That is, imitability does not guarantee identifiability of the imitators reward distribution. In fact, Zhang et al. further show that if an expert and imitator share the policy space (possible actions), the the policy itself is always imitable. When the policy spaces do not agree, the criteria become more complicated. We now explore some graphical criteria for this case.

Definition 4.29 (Imitation Backdoor [49])**.** *Consider a causal system with diagram G and policy space Π. We say $\boldsymbol{Z}$ satisfies the imitation backdoor criterion (i-backdoor) with respect to $\langle G, \Pi\rangle$ if and only if the Z is a subset of the parents of Π and Y is conditionally independent of X given $\boldsymbol{Z}$ in the graph with edges out of X removed. Formally, $\boldsymbol{Z} \subseteq Pa(\Pi)$ and $(Y \perp\!\!\!\perp X \mid \boldsymbol{Z})_{G_{\underline{X}}}$.*

For an example of how this definition applies, consider figure 19 (a) with set $\{Z\}$. Here we of course have $\{Z\}$ is in the parents of the policy space (inclusive of the space itself). Also, removing edge $X \to Y$ and conditioning on Z still leaves path $X \leftarrow L \to Y$ as a 'backdoor'. In figure 19 (b), however, the $L \to Y$ edge is removed and we no longer have an i-backdoor set.

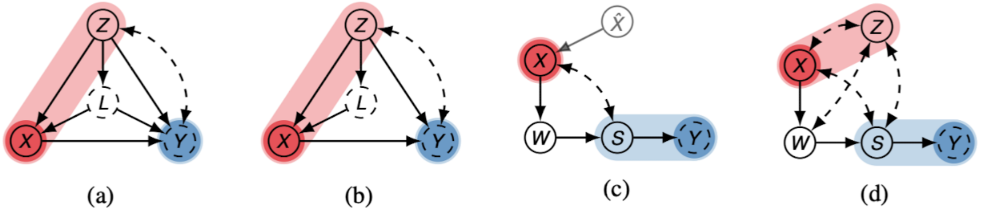


Figure 19: Causal diagrams showing examples where imitation learning can or cannot occur. Blue variables indicate latent reward variable, while red variables represent action. Light red indicates the inputs to the policy space, and light blue represents the minimal imitation surrogates. Figure extracted from [49].

The i-backdoor criterion is used to characterise when imitation of an expert is possible given that the reward variable is unobserved.

Theorem 4.16 (Imitation by Backdoor [49]). *Given a causal diagram G with policy space Π, we say that the distribution $P(y)$ is imitable with respect to $\langle G, \Pi \rangle$ if and only if there exists some i-backdoor admissible set $\boldsymbol{Z}$ in this causal system. Further, the policy itself can be determined and is given by $\pi(x \mid pa(\Pi)) = P(x \mid \boldsymbol{z})$.*

Applying theorem (4.16) we can see that using set $\{Z\}$ in fig 19 (b), an imitating policy is learnable. Further, the policy itself is given by $\pi(x \mid z) = P(x \mid z)$. These results are impressive, but Zhang et al. further point out that the i-backdoor requirement is not necessary for imitation of expert performance. Consider figure 19 (c) in which variable S mediates all actions (intervention on X) on the outcome (latent reward Y). We could imagine that learning a distribution over S could be sufficient for imitation of the distribution of Y in this case. This train of thought motivates the following definitions and formalism's.

Definition 4.30 (Imitation Surrogate [41]). *Consider a causal diagram G with policy space Π. Take $\boldsymbol{S}$ as an arbitrary subset of the observations, $\boldsymbol{O}$. We say $\boldsymbol{S}$ is an imitation surrogate (i-surrogate) with respect to $\langle G, \Pi \rangle$ if $(Y \perp\!\!\!\perp \hat{X} \mid \boldsymbol{S})_{G \cup \Pi}$. Here $G \cup \Pi$ is the graph obtained by adding directed edges from $Pa(\Pi)$ to X. $\hat{X}$ is a new parent of X that has been added in this procedure.*

We say that the imitation surrogate is minimal if there is no subset of it such that the subset is also an imitation surrogate. A simple example of this is visible in figure 19 (c) where both $\{W, S\}$ and $\{S\}$ are surrogates, but $\{S\}$ is the minimal surrogate. Figure 19 (d) poses an additional problem in that the addition of the collider Z to (c) means $P(s \mid do(\pi))$ is not identifiable even though we have an i-surrogate S. The authors tackle this problem by realising that having a subspace of the policy space that yields an identifiable distribution is sufficient to solve the imitation learning task. Without delving into the details here, the authors of [49] implement a confounding robust imitation learning algorithm and apply it to several interesting problems in which the causal approach is shown to be superior to naive imitation learning approaches.

This completes the discussion about this task. I expected causal imitation learning to become an active area of research as this paper was only recently released, being the first of its kind. This also concludes the development of the theory necessary for engaging with state-of-the-art research in causal reinforcement learning.

5 Conclusions

This paper briefly introduced necessary notions of causal inference and reinforcement learning before delving into state-of-the-art work at the overlap between these two fields. This work was introduced under the guise of causal reinforcement learning (CRL) with the promise of tackling six difficult tasks in causal inference and reinforcement learning by combining theory from the different fields. A range of specific modern work was surveyed and key results were developed so as to coherently develop a comprehensive overview of the emerging field. In the area of generalised policy learning specific successes in areas of dynamic treatment regimes were discussed. This has promising ramifications for optimal decision making in healthcare, for example. Next, we discussed and developed some graphical theory and criteria for selecting (possibly) optimal interventions in a causal system. We then proceeded to discuss counterfactual decision making in which we extended traditional reinforcement learning formalism's and experimental design to account for unobserved confounding. The next task involved generalisation of results across domains. We developed necessary theory and graphical criteria for discovering possibility of (generalised) *identifiablity* from a select set of variables in a causal domain. Next, we discussed modern advancements in causal structure learning and presented some impressive results. Finally, we briefly discussed some applications of earlier theory to imitation learning in a causal domain.

6 Discussion & Further Investigation

This paper was ambitious in its scope and limitations, to say the least. The rate at which development is occurring in both fields of reinforcement learning and causal inference is staggering - certainly too fast to write an always-relevant, detailed and complete introduction to CRL. Trying to combine these fields is a behemoth of a task, and has only in very small part (I hope) been achieved in this paper. The primary focus of this paper was developing the theory necessary for understanding and dealing with modern research in the emerging, inter-disciplinary area of causal reinforcement learning. There was limited room for detailed expansion throughout the paper. It is my hope that, despite the brevity at times, the content presented is digestible. If there's one thing the brevity does leave, it is the immense scope for future investigation and research.

Of primary interest to me is in how to augment current methods of reinforcement learning to work towards a more generally applicable and intelligent agents. From my perspective, the most naive behaviour of an RL agent is its lack of awareness of causal relationships. Of course, the interventional behaviour of an RL agent results in cause-effect results, but most RL agents are naive to the underlying data generating structure. In noisy, ever-changing, real-world domains, agents require the ability to identify a changing landscape, as well as transfer and apply their previous knowledge efficiently and accurately. They should also be able to identify the reasons for their actions and explain why they have decided to take certain actions. All of this becomes much clearer with an explicit causal model of the world. Transfer of causal knowledge becomes even more important when we start to consider multi-agent and imitation learning scenarios. We are already seeing many impressive CRL theoretical results appear in the literature. Perhaps the most exciting thing is how these will be applied to form groundbreaking applications. I look forward to future developments in this area. If not for interesting applications sake, the field offers a wonderful mix of graph theoretic, complexity theory, statistics and reinforcement learning results (among others).

Acknowledgements

I would like to thank Dr Jonathan Shock for so eagerly guiding and freely allowing me to explore almost any area of interest - from neuroscience and reinforcement learning to causal inference. Additionally, I would like to thank Jeremy du Plessis for the many discussions, teamwork, and help, despite my naivete with regards to reinforcement learning. It wasn't so bad after all! Finally, I would like to thank Sanghack Lee for taking the time to respond to my queries by email.

Plagiarism Declaration

1. I know that plagiarism is a serious form of academic dishonesty.
2. I have read the UCT document Avoiding Plagiarism: A guide for students, am familiar with its contents and have avoided all forms of plagiarism mentioned there.
3. Where I have used the words of others, I have indicated this by the use of quotation marks.
4. I have referenced all quotations and properly acknowledged other ideas borrowed from others.
5. I have not and shall not allow others to plagiarise my work.
6. I declare that this is my own work.

Signature:

References

[1] David Hume. *An enquiry concerning human understanding.* Hackett, 2 edition, 1993.

[2] Ronald A. Fisher. Cancer and smoking. *Nature*, 182:596–596, 1958.

[3] Sir Ronald A. Fisher. Smoking: The cancer controversy. 1960.

[4] L. Penrose. Cancer and smoking. *Nature*, 182:1178–1178, 1958.

[5] D. Rubin. Causal inference using potential outcomes. *Journal of the American Statistical Association*, 100:322 – 331, 2005.

[6] Andrew Gelman et al. Resolving disputes between j. pearl and d. rubin on causal inference. https://statmodeling.stat.columbia.edu/2009/07/05/disputes_about/, 2009.

[7] Judea Pearl and Dana Mackenzie. *The Book of Why*. Basic Books, New York, 2018.

[8] Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. On pearl's hierarchy andthe foundations of causalinference. unpublished, 2020.

[9] Elias Bareinboim. Causal reinforcement learning. ICML 2020, 2020.

[10] J. Peters, D. Janzing, and B. Schölkopf. Elements of causal inference: Foundations and learning algorithms. 2017.

[11] Elias Bareinboim and J. Pearl. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113:7345 – 7352, 2016.

[12] Richard Sutton and Andrew Barto. *Reinforcement Learning: An Introduction*. MIT Press, second edition edition, 2018.

[13] Sergey Levine. Deep rl at berkeley: Cs285. http://rail.eecs.berkeley.edu/deeprlcourse/, 2019.

[14] Junzhe Zhang and Elias Bareinboim. Transfer learning in multi-armed bandits: A causal approach. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 1340–1346, 2017.

[15] J. Langford and T. Zhang. The epoch-greedy algorithm for contextual multi-armed bandits. In *NIPS 2007*, 2007.

[16] Aurélien Garivier and O. Cappé. The kl-ucb algorithm for bounded stochastic bandits and beyond. *ArXiv*, abs/1102.2490, 2011.

[17] J. Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. In *NeurIPS*, 2019.

[18] J. Zhang and Elias Bareinboim. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *ICML 2020*, 2020.

[19] Hongseok Namkoong, Ramtin Keramati, S. Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *ArXiv*, abs/2003.05623, 2020.

[20] J. Zhang and Elias Bareinboim. Bounding causal effects on continuous outcomes. 2020.

[21] Elias Bareinboim, Andrew Forney, and J. Pearl. Bandits with unobserved confounders: A causal approach. In *NIPS*, 2015.

[22] Sanghack Lee and Elias Bareinboim. Structural causal bandits: Where to intervene? In *NeurIPS*, 2018.

[23] Sanghack Lee and Elias Bareinboim. Structural causal bandits with non-manipulable variables. In *AAAI*, 2019.

[24] S. Lee. Characterizing optimal mixed policies: Where to intervene and what to observe. 2020.

[25] P. Auer, Nicolò Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47:235–256, 2004.

[26] Andrew Forney and Elias Bareinboim. Counterfactual randomization: Rescuing experimental studies from obscured confounding. In *AAAI*, 2019.

[27] W. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25:285–294, 1933.

[28] J. Zhang and Elias Bareinboim. Markov decision processes with unobserved confounders : A causal approach. 2016.

[29] I. Szita and Csaba Szepesvari. Model-based reinforcement learning with nearly tight exploration complexity bounds. In *ICML*, 2010.

[30] Andrew Forney, J. Pearl, and Elias Bareinboim. Counterfactual data-fusion for online reinforcement learners. In *ICML*, 2017.

[31] J. Heckman. Randomization and social policy evaluation. 1991.

[32] R.A. Fisher. *The Design of Experiments.* The Design of Experiments. Oliver and Boyd, 1935.

[33] W. W. Stead, Starmer J. M., and M. McClellan. Beyond expert based practice. *Evidence-Based Medicine and the Changing Nature of Healthcare: 2007 IOM Annual Meeting Summary*, 2008.

[34] J. Zhang and Elias Bareinboim. Can humans be out of the loop? 2020.

[35] Elias Bareinboim and J. Pearl. Transportability from multiple environments with limited experiments: Completeness results. In *NIPS*, 2014.

[36] S. Lee, Juan David Correa, and Elias Bareinboim. General identifiability with arbitrary surrogate experiments. In *UAI*, 2019.

[37] Jin Tian and J. Pearl. Studies in causal reasoning and learning. 2002.

[38] Sanghack Lee. General identifiabilitywith arbitrary surrogate experiments. AAAI 2020, presented at UAI 2019, 2019.

[39] C. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37:424–438, 1969.

[40] P. Spirtes, C. Glymour, and R. Scheines. Causation, prediction, and search, second edition. In *Adaptive computation and machine learning*, 2000.

[41] M. Kocaoglu, Karthikeyan Shanmugam, and Elias Bareinboim. Experimental design for learning causal graphs with latent variables. In *NIPS*, 2017.

[42] Julien Bensmail, Marthe Bonamy, and Hervé Hocquard. Strong edge coloring sparse graphs. *Electron. Notes Discret. Math.*, 49:773–778, 2015.

[43] A. Ghassami, Saber Salehkaleybar, N. Kiyavash, and Elias Bareinboim. Budgeted experiment design for causal structure learning. In *ICML*, 2018.

[44] Christopher Meek. Causal inference and causal explanation with background knowledge. *ArXiv*, abs/1302.4972, 1995.

[45] G. Nemhauser, L. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14:265–294, 1978.

[46] M. Kocaoglu, A. Jaber, Karthikeyan Shanmugam, and Elias Bareinboim. Characterization and learning of causal graphs with latent variables from soft interventions. In *NeurIPS*, 2019.

[47] A. Jaber and M. Kocaoglu. Causal discovery from soft interventions with unknown targets: Characterization and learning. 2020.

[48] P. Abbeel and A. Ng. Apprenticeship learning via inverse reinforcement learning. In *ICML '04*, 2004.

[49] J. Zhang, Daniel Kumor, and Elias Bareinboim. Causal imitation learning with unobserved confounders. In *June preprint*, 2020.

[50] Karl Tuyls, Julien Pérolat, Marc Lanctot, Joel Z. Leibo, and Thore Graepel. A generalised method for empirical game theoretic analysis. In *AAMAS*, 2018.

[51] Petter N. Kolm and Gordon Ritter. Modern perspectives on reinforcement learning in finance. *Econometrics: Mathematical Methods Programming eJournal*, 2019.

[52] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 2016.

[53] Massimo Silvetti and Tom Verguts. Reinforcement learning, high-level cognition, and the human brain. 2012.

[54] Tingwu Wang, Xuchan Bao, Ignasi Clavera, Jerrick Hoang, Yeming Wen, Eric Langlois, S. Zhang, Guodong Zhang, Pieter Abbeel, and Jimmy Ba. Benchmarking model-based reinforcement learning. *ArXiv*, abs/1907.02057, 2019.

[55] Maor Gaon and Ronen I. Brafman. Reinforcement learning with non-markovian rewards. In *AAAI*, 2020.

7 Appendices

7.1 Appendix A: Causal Inference

In this section we briefly discuss some of the fundamentals of causal inference and discuss some of the motivating examples for the development of graphical causal theory. Much of this development is based on work by Jonas Peters [10]. This work was similarly written up for my personal blog while researching this section. We develop notions for models of two or three variables for ease of understanding. These notions do generalise to higher dimensions under the guise of multivariate causal models. Multivariate models require some additional concepts, but these are discussed as required in the main text body since the multivariate case is the one we will work with throughout the CRL literature.

7.1.1 Theory

Perhaps the most fundamental assumption in causal inference is that of a common cause.

Principle 7.1 (Reichenbach's common cause)**.** *Given two statistically dependent random variables X and Y, there exists a third variable Z that causally influences both X and Y - a common cause. A special case is, of course, the situation when Z is X or Y. Also, this common cause 'shields' the two variables from each other. Formally, given Z, X is independent of Y.*

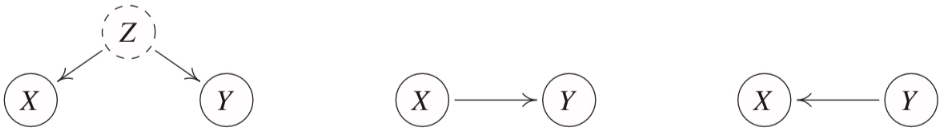


Figure 20: Figure showing Reichenbach's common cause principle. Here we see the three possibilities - Z is a third variable, Z is X, and Z is Y. Extracted from [10].

It should now be clear that there is a very important difference between conditioning on a variable and intervening on that variable. By conditioning we are 'asking' our data what we *observe* when one variable has a certain value. When intervening, however, we are *changing* a variable and seeing how the system responds. In other words, we are changing a variable without it changing the other variables it is associated with. This is where the modelling of dependencies becomes crucial. We are now ready to discuss Pearl's causal hierarchy [7].

7.1.2 Structural Causal Models

Jonas Peters introduces SCMs using the classic problem in machine learning of recognising and classifying handwritten digits - usually done using the MNIST dataset. Let us consider how this data was collected. Let's say in scenario (i) we tell someone the number to write. In other words, we give them the 'label' and they write the corresponding number. In scenario (ii), the person decided the number to write and gives the digit an associated label. The resulting dataset will look exactly the same, and so the joint density distribution will be exactly the same. However, we know there is a fundamental difference. Calling the label Y and the handwritten digit X, if we intervene on Y by changing the label, the result will be a different handwritten digit. In some sense, the label we 'give' causes the resulting handwritten digit. In scenario (ii), changing the label does not change the handwritten digit. There is a common cause - the intention of the person writing - but Y does not cause X. SCMs are formally discussed in the preliminaries section of the main text.

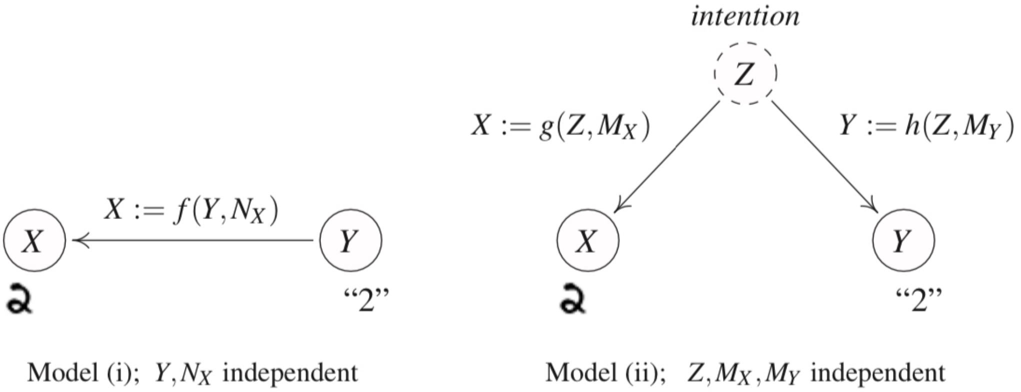

$X := f(Y, N_X)$
X
Y
"2"
Model (i); Y, N_X independent
intention
Z
$X := g(Z, M_X)$
$Y := h(Z, M_Y)$
X
Y
"2"
Model (ii); Z, M_X, M_Y independent

Figure 21: Figure showing the possibilities for the causal mechanism of collecting the handwritten MNIST dataset. In this case both models induce equivalent joint densities over the variables of interest. Intervening in the systems produce different results. Extracted from [10].

7.1.3 Assumptions for Causal Inference

The assumptions in the theory of causal inference are crucial to consider as they determine what and how we can learn causal models from data. Assumptions of independence are particularly important in this context.

If $X \to Y$ is some causal structure, then:

1. It is possible to intervene locally on the variable X without changing Y. This is the notion of independence.
2. Conditioning on a common cause leads to independence. In this scenario, $p(x)$ and $p(y|x)$ are autonomous, modular or invariant quantities.

Principle 7.2 (Independent mechanisms). *The causal generative process of system's variables is composed of autonomous modules that do not inform or influence each other.*

This principle makes sense if we think of these *modules* as being physical mechanisms in the real world. The idea is that if we intervene by changing one mechanism, the other will be unaffected. This assumption can help with the idea of transferring knowledge between related domains. How different is a robot serving ice cream from an industrial robot moving car parts? Crucially, all this is to say that the mechanism generating the effect from its cause contains no information about the mechanism generating the cause.

7.1.4 Learning Causal Models

Previously, we discussed the idea that an SCM induces a joint distribution over the variables of interest. For example, the SCM $C \to E$ induces $P_{C,E}$. Naturally, we wonder whether we can *identify*, in general, whether the joint distribution came from the SCM $C \to E$ or $E \to C$. It turns out, the graphs are not unique. In other words, structure is not *identifiable* from the joint distribution. Another way of phrasing this is the graph adds an additional layer of knowledge.

Proposition 7.1 (Non-uniqueness of graph structures). *For every joint distribution $P_{X,Y}$ of two real-valued variables, there is an SCM*

$$Y = f_Y(X, N_Y), X \perp Y,$$

where f_Y is a measurable function and N_Y is a real-valued noise variable.

This proposition indicates that we can construct an SCM from a joint distribution in any direction. This is crucial to keep in mind, especially if we plan on trying to use observational data to infer causal structure. We are now ready to discuss some methods of identifying cause and effect with some *a priori* restrictions on the class of models we are using.

Additive Noise Models Our first class of model are the linear non-Gaussian acyclic models (LiNGAMs). Here we assume that the effect is a linear function of the cause up to some additive noise term,

$$E = \alpha C + N_E, \quad \alpha \in \mathbb{R}$$

Note, we are explicitly removing the possibility that the additive noise is Gaussian in nature. With this restrictive assumption in place, we can formulate the identifiability result we are looking for:

Theorem 7.1 (Identifiability of linear non-Gaussian models). *Given the joint distribution $P_{X,Y}$ having linear model*

$$Y = \alpha X + N_Y, \quad N_Y \perp X,$$

with continuous random variables X, N_Y, Y, there exists $\beta \in \mathbb{R}$ and random variable N_X such that

$$X = \beta Y + N_X, \quad N_X \perp Y,$$

if and only if N_Y and X are Gaussian.

Peters provides a lucid example of this in action. Here we have uniform noise over X and a linear relationship between the variables of interest. The backwards model, shown in blue, is not valid because the noise term over X is not independent of the variable Y, violating the independence condition.

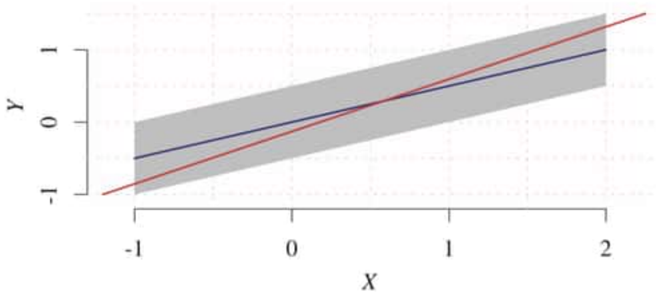


Figure 22: Figure [10] showing uniform noise over X and a linear relationship between the variables of interest. The backwards model, shown in blue, is not valid because the noise term over X is not independent of the variable Y, violating the independence condition

Interestingly enough, non-Gaussian additive noise can be applied to estimating the arrow of time from data! We are now ready to extend the discussion the nonlinear additive noise models.

Definition 7.1 (Additive noise model (ANM))**.** *The joint distribution $P_{X,Y}$ is said to admit an ANM from X to Y if there is a measurable function f_Y and a noise variable N_Y such that*

$$Y = F_Y(X) + N_Y, \quad N_Y \perp X.$$

We can extend the identifiability condition to include ANMs as well. For brevity I shall not include it. Peters, however, provides a nice description and proof sketch in chapter 4 of his book. Further extensions, such as to discrete ANMs and post-nonlinear models, are possible and described in the literature.

Structure Identification We have formulated causal relationships as directed acyclic graphs (DAGs) up to this point. The problem we now face is how to apply the identifiability results from previous discussions to identifying and building causal structural models for the data. This is a very challenging learning problem and the results we now discuss are very much current research and bound to change.

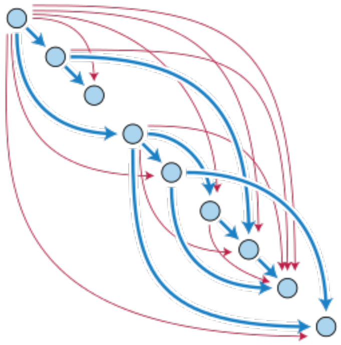


Figure 23: Example of a DAG with possible relationships between variables. This indicates the difficulty of the nature of the learning problem as we scale the size of the causal model. Extracted form [10].

Once again we can look at the simple case of additive noise models. The first method tests the *independence of the residuals* in the data. This is a special case of the RESIT algorithm we shall discuss later. This algorithm runs as follows:

1. Regress Y on X such that we can write Y as a function of X and a noise term.
2. Test the independence between $Y - \hat{f}_Y(X)$ and X.
3. Repeat steps (1) and (2), switching the roles of X and Y.
4. If one direction is independent but the other isn't, infer the former direction as causal.

Peter's provides an informative example of this testing procedure in action using a simple example. The left side corresponds to the X data while the right displays the Y data. Regression procedures are shown in the top row with corresponding residuals shown below. It is clear that there is a dependence for the X on Y direction. Applying the algorithms above, we infer a causal direction, $X \to Y$.

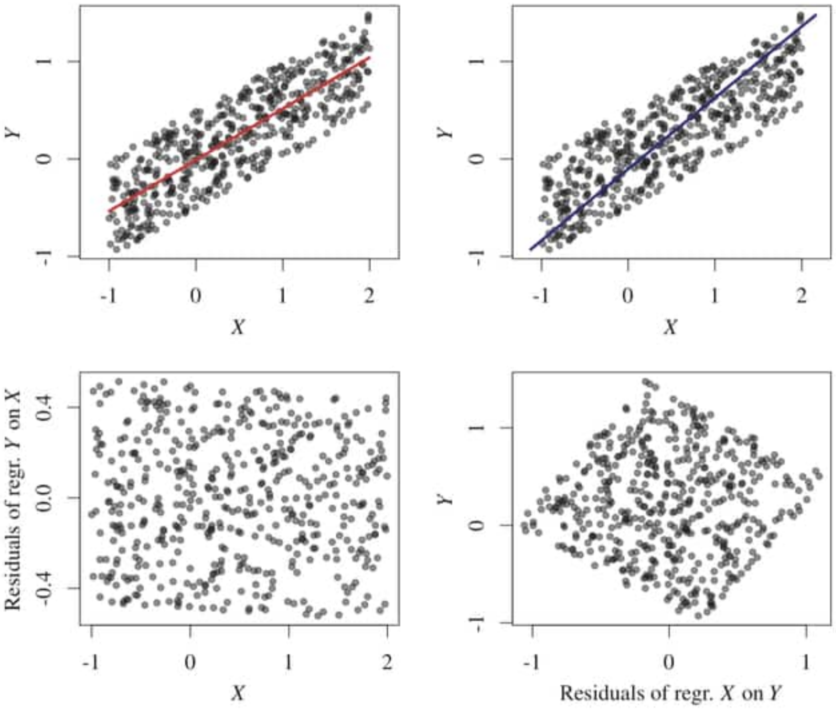


Figure 24: Example of applying causal structure learning algorithm to simple linear scenarios. First we have a classic linear relationship in the top-left. Top-right shows an attempt at reverse regression which shows dependence on residuals. The bottom graphs show the residuals in each case. Clearly there is a dependence on the right hand side. Example from [10].

An alternative approach is to apply a maximum likelihood method. We compare the causal directions by comparing the associated likelihood scores. These are computed as

$$L_{X\to Y} = -\log \hat{var}[X] - \log \hat{var}[R_Y],$$

where R_Y are the residuals. We repeat this for the $Y \to X$ direction.

Semi-supervised Learning Suppose we are given some training data with labels in the form $(D_1, A_1), \dots, (D_n, A_n)$ where each of the D_i is a dataset such that

$$D_i = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$$

containing i.i.d. realisations from the distribution $P^i_{X,Y}$. The label $A_i \in \{\to, \leftarrow\}$ such that it describes the causal direction between the pair (X, Y). Since we now have labled training data, this learning problem becomes a supervised statistical learning prediction problem. In supervised learning scenarios we are given a sample of labled data points drawn from some joint distribution. Formally,

$$(X_1, Y_1), \dots, (X_n, Y_n) \sim^{\text{i.i.d.}} P_{X,Y}.$$

This gives us information about the conditional distribution $P_{X|Y}$ which we can exploit. In semi-supervised learning we are given an additional m unlabled data points. Since these points are not conditioned, one would hope this would give information about the unconditional distribution, P_X. In previous discussions we developed notions for talking about causal structure of two variables. In this case, a variable is either *causal* or *anti-causal.* Traditionally, machine learning engineers do not consider the underlying causal structure. This is a danger. Contrary to popular belief, more data does not always solve our issues - in fact, it can make it worse! Recall, we discussed that causal conditional distributions are independent of each other. This necessarily means the distribution of the one does not contain information about the distribution of the other. The following figure nicely demonstrates the performance of self-supervised learning agents on different datasets. We immediately notice that when causal data is used to predict an effect, additional data *hurts* performance when comparing against some base classifier. This is remarkable, albeit intuitive.

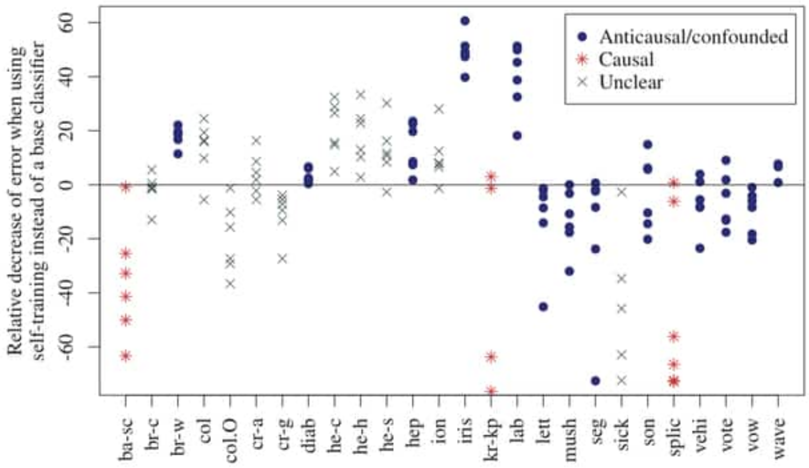


Figure 25: Motivating example for acknowledging causal relationships for statistical learning purposes. Here we notice that applying semi-supervised learning techniques to anti-causal data can actually lead to a decrease in performance. This result is extracted from [10].

7.2 Appendix B: Reinforcement Learning

The following section is largely sourced from work conducted for a reading module in reinforcement learning. The field of reinforcement learning (RL) is at a crossroads between optimal control, animal psychology, artificial intelligence and many other technical fields including game theory [50] and finance [51]. It has seen a surge of interest with high profile successes, such as the dominance of AlphaGo [52] over 18 time world Go champion, Lee Sedol. The idea of reinforcement learning emerged out of the study of animal psychology and the concept of trial-and-error learning. Edward Thorndike coined the *Law of Effect* in 1898, introducing the idea of reinforcement in the context of learning. He proposed that if a stimulus is followed by some successful response (reward in RL), then that stimulus-response connection will be strengthened [53]. Recent successes of the theory of reinforcement learning at solving complex computational tasks have seen a resurgence in the field.

Reinforcement learning can be seen as a subset of machine learning alongside supervised and unsupervised methods. Whereas supervised and unsupervised learning deal with labelled and unlabelled static data respectively, reinforcement learning deals with how an agent should learn to make decisions in an abstract environment with a potentially changing landscape. More recently, the field has been subdivided into unsupervised RL and supervised RL. In supervised RL we explicitly program the reward structure from which the agent learns. In unsupervised RL the agent can learn from intrinsic motivations, such as curiosity.

RL is, for the most part, a formalism for the idea of trial-and-error learning. At each time-step an agent in an environment with a well defined state can perform some action. This action triggers a response from the environment in which the environment rewards the agent for taking this action and moves the agent to a new state. Naturally, the goal of the agent is to maximise the expected reward over time - the return. Further complexity arises in that the agent does not explicitly know how the environment will react and may only have partial information about what state it is currently in. For example, a human playing Montezuma's Revenge (an Atari game used for benchmarking RL agents) can only see a portion of the environment and cannot be sure about what exact state the surrounding environment is in. It must attempt to make optimal decisions with imperfect information.

Reinforcement learning algorithms are usually categorised as falling into one of two categories - model-free reinforcement learning (MFRL) or model-based reinforcement learning (MBRL). In MFRL an agent directly learns a value function or policy for interacting in the environment by observing rewards and state transitions, while in MBRL the agent uses its interactions with the environment to learn about the environment's dynamics, and thereby model it. Model-free methods have enjoyed success in areas such as robotics and computer games, however they are plagued by high sample inefficiencies [54]. These sample inefficiencies often limit the application of these methods to simulated environments which are less complex than real world dynamics. Learning a model of the environment allows an agent to significantly reduce the dimensionality and complexity of the environment it interacts with, allowing for much improved sample efficiency and performance in many simulated and real world scenarios. By learning a model of the environment, an agent can apply well known supervised learning techniques to plan an optimal strategy for maximising reward. Learning a good model of the environment is a non-trivial task as modelling errors can prove crippling to many tasks. This is the problem of model-bias.

7.2.1 The Reinforcement Learning Problem

The problem of reinforcement learning (RL) is that of how to map states to actions so as to maximise some numerically encoded reward signal [12]. As discussed in the introduction, the reinforcement learning agent experiences a changing landscape based on the actions it takes and attempts to learn an optimal sequence of actions by applying the concept of trial-and-error with reinforcement theory. This problem can be formally stated in terms of Markov decision processes (MDP) within what Sutton and Barto [12] refer to as the agent-environment interface. MDPs are an appropriate formalism for the RL problem since many (especially classical) reinforcement learning algorithms make use of dynamic programming techniques where the exact mathematical model of the problem is available. In some sense, reinforcement learning is the study of approximate dynamic programming.

The agent-environment interface describes how an agent and the environment the agent is interacting with communicate and interplay. It is cyclical in nature, as shown in figure 26.

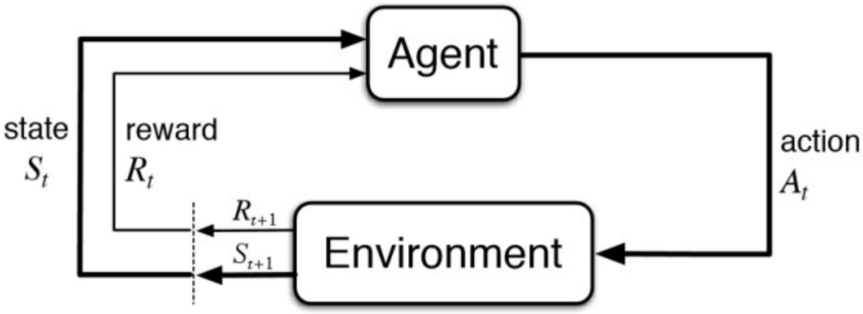


Figure 26: Diagram showing the flow of components in an MDP. An agent selects action A_t. The environment returns the next state and an associated reward to the agent, dependent on the selected action. The process then repeats [12].

We can begin to formalise the RL problem by first noting that the reward an agent receives is only dependent on the current state it is in, as well as the action it decides to take. In other words, the history of how the agent got to the current state is irrelevant. In statistical terms, we can say it has the Markov property. Of course, we can formulate problems in which the environment has some memory of the agent's decisions [55], but this is not the common approach in RL and shall not be discussed further.

Definition 7.2. *A state has the Markov property if for t = 1, 2, ..., $P(X_{t+1} = i_{t+1}|X_t = i_t, \dots, X_1 = i_1, X_0 = i_0) = P(X_{t+1} = i_{t+1}|X_t = i_t)$.*

The agent exists in an environment of states which all have the Markov property. The states thus form a Markov chain.

Definition 7.3. *A discrete-time stochastic process is a Markov chain if every state has the Markov property. Formally, it is a tuple (S, P) of states S and transition probabilities P.*

Further, since moving from one state to another results in a reward, we can define a Markov reward process.

Definition 7.4. *A Markov reward process (MRP) is a Markov chain where rewards are associated with each state. Formally, it is a 4-tuple (S, P, R, γ) of states S, transition probabilities P, rewards R, and a discount factor γ.*

Further, the agent exists in the MRP and can make decisions. This leads to the definition of a Markov decision process (MDP) with which most RL problems can be formulated.

Definition 7.5. *A Markov decision process (MDP) is a Markov reward process with actions that can be taken. Formally, it is a 5-tuple (S, A, P, R, γ) of states S, actions A, transition probabilities P, rewards R, and a discount factor γ.*

In many modern problems an agent can only observe part or some representation of states in the state space. This requires the further abstraction of a partially observed Markov decision process (POMDP).

Definition 7.6. *A partially observed Markov decision process (POMDP) is a Markov decision process in which the agent cannot observe the underlying state. Instead, there are observations and associated observation probabilities which allow inference about the underlying state. Formally, it is a 7-tuple $(S, A, P, R, \Omega, O, \gamma)$ of states S, actions A, transition probabilities P, rewards R, observations Ω, observation probabilities O, and a discount factor γ.*

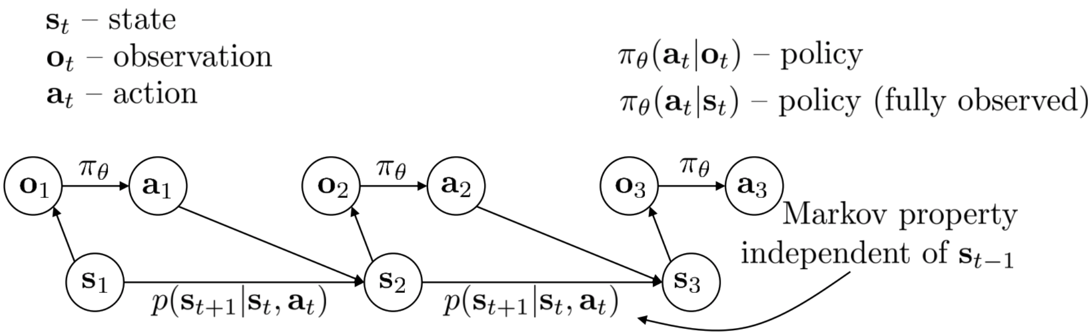

$\mathbf{s}_t$ – state
$\mathbf{o}_t$ – observation
$\mathbf{a}_t$ – action
$\pi_\theta(\mathbf{a}_t|\mathbf{o}_t)$ – policy
$\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ – policy (fully observed)
$\mathbf{o}_1$
π_θ
$\mathbf{a}_1$
$\mathbf{o}_2$
π_θ
$\mathbf{a}_2$
$\mathbf{o}_3$
π_θ
$\mathbf{a}_3$
Markov property
independent of $\mathbf{s}_{t-1}$
$\mathbf{s}_1$
$p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
$\mathbf{s}_2$
$p(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$
$\mathbf{s}_3$

Figure 27: Figure showing a diagram of relationships between variables in a POMDP. Figure extracted from [13].

7.2.2 Exploration and Exploitation

One of the inherent problems an agent faces in some arbitrary environment is how to decide whether to explore and discover more of the world around it, or to rather apply what it already knows and exploit this knowledge to maximise the expected return. There is a trade-off since exploiting what we already know will likely lead to some amount of reward, but with imperfect information about the world around the agent there is a possibility that the agent is missing out on some large reward it does not yet know about. There is thus a problem of ensuring good exploration of the state-space while also ensuring at some point we apply what we have learned to ensure good reward. This problem becomes even harder when the environments dynamics are themselves a function of time, and so what we have learned in previous time-steps may become invalid as we progress. This is an active area of research and some possible solutions are presented and discussed in the context of model-based reinforcement learning in later sections.

Multi-armed bandit problems consist of of choosing among multiple discrete arms, each yielding some specific reward with some unknown probability. The common example is that of a slot machine. The reward function of the bandit is unknown, so an agent must interact with it to *learn* and approximate its reward function. At each point it must make the decision of whether to learn more about the machine or to exploit the knowledge it has gained.