AIStats Submission Sept 26: up to "CBNs are causal theories"

October 8, 2019

1 Introduction

It is widely accepted that causal knowledge and statistical knowledge are distinct. Statistics is concerned with *association* while causation is concerned with *consequences*; a distinction of this type goes back at least to Hume (Morris and Brown, 2019). There are a number of modern approaches to representing causal knowledge, which broadly fall into two categories: those based on some means of representing counterfactual relationships such as *potential outcomes* (Rubin, 2005) and those based on causal graphical models such as *causal Bayesian networks* (Pearl, 2009). There have been a number of attempts to bring these two views together Richardson and Robins (2013); Shpitser (2008).

Perhaps as a result of the requirements of representing causal knowledge, both approaches feature idiosyncratic elements, and it isn't fully transparent where they follow statistical theory and where they depart from it. For example, causal Bayesian networks feature *do* operations that appear to have no statistical counterpart, while counterfactual problems can be formulated as missing data problems. However, Pearl (2009) claims that the counterfactual approach subsumes the Bayesian network approach, suggesting that perhaps *do* operations are statistical objects in disguise or, alternatively, that counterfactual random variables are somehow non-standard.

We develop an approach to causal inference taking statistical decision theory (Wald, 1950) as a starting point and, with conceptual cues from Savage (1972), introducing the notion of utilities and consequences. As a result our approach has very clear connections with statistical theory. We develop causal statistical decsion theory (CSDT) which features *causal theories* as the central object of study. Causal theories play a role analogous to *statistical experiments* in ordinary statistical decision theory. They differ in that, where a statistical experiment gives you a probability measure, a causal theory gives a stochastic map.

Causal Bayesian networks themselves have a natural representation as causal theories. Against Pearl's claim that counterfactual models subsume interventional ones, we find that some arbitrary choices must be made in order to represent potential outcomes models as causal theories. Nonetheless, we propose a plausible strategy for representing potential outcomes models as causal theories. We find that these two approaches yield very different causal theories from one another, though these differences may disappear when we move from "idealised" theories to more realistic theories that we might actually use to make decisions.

We then turn our attention to a notable feature of the idealised causal theories induced by both approaches: they each allow for unrealistically fine control over many of their consequences. We posit that these theories are intended to inform realistic theories used for making decisions. We show that this is possible if the realistic theories can be derived from the idealised theories via what we term *coarsening*. To our knowledge, this is the first attempt to formalise the notion of "stable knowledge" that is often raised as a key potential advantage of causal understanding over purely statistical learning (Arjovsky et al., 2019; Pearl, 2009; Rubin, 2005).

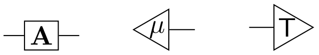
2 Definitions & Notation

We use the following standard notation: $[n]$ refers to the set of natural numbers $\{1, ..., n\}$. Sets are ordinary capital letters X, σ-algebras are script letters $\mathcal{X}$ while random variables are sans serif capitals $\mathsf{X} : \_ \to X$. All sets mentioned are understood to be equipped with measures. The calligraphic $\mathcal{G}$ refers to a directed acyclic graph rather than a σ-algebra. Probability measures are greek letters μ, ξ, γ and stochastic maps are bold capitals $\mathbf{C}, \mathbf{H}$. Sets of probability measures or stochastic maps are script capitals: $\mathcal{H}$, $\mathcal{T}$, $\mathcal{J}$. We write the set of all probability measures on $(X, \mathcal{X})$ as $\Delta(\mathcal{X})$. $\delta_x : (X) \to [0, 1]$ is the probability measure such that $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise.

A stochastic map or Markov kernel is a map $\mathbf{A} : X \to \Delta(\mathcal{Y})$. We write the first argument of a Markov kernel as a subscript; for $x \in X$, $G \in \mathcal{Y}$, $\mathbf{A}_x$ is a probability measure on X and $A_x(G) \in [0, 1]$ is the measure of G. For $\mathbf{A}$ to be a Markov kernel we also require that the function $x \mapsto A_x(G)$ must be measurable for all $G \in \mathcal{Y}$. For $\mathbf{C} : X \times Y \to \Delta(\mathcal{Z})$ and $x \in X$ we will write $\mathbf{C}_x$ for the "curried" map $y \mapsto \mathbf{C}_{x,y}$.

We can use a notation similar to matrix-vector products to represent relationships with Markov kernels. Probability measures $\mu \in \Delta(\mathcal{X})$ can be read as row vectors, Markov kernels as matrices and measurable functions $\mathsf{T} : Y \to T$ as column vectors. Defining $\mathbf{B} : Y \to \Delta(\mathcal{Z})$ we have $\mu\mathbf{A}(G) = \int \mathbf{A}_x(G)d\mu(x)$, $\mathbf{A}\mathbf{B}_x(H) = \int \mathbf{B}_y(H)d\mathbf{A}_x(y)$ and $\mathbf{A}\mathsf{T}(x) = \int \mathsf{T}(y)d\mathbf{A}_x(y)$. The tensor product is $(mathbfA \otimes \mathbf{B})_{x,y}(G,H) = \mathbf{A}_x(G)\mathbf{B}_y(H)$ where the product on the left is scalar multiplication. Kernel products are associative and the product of kernels is always a kernel itself (Çinlar, 2011).

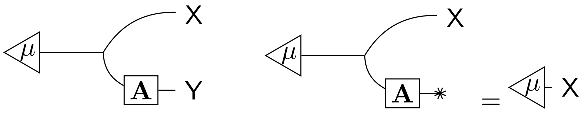
Some elaborate constructions are unwieldy in inline product notation. Here we use string diagrams. String diagrams can always be interpreted as a mixture of kernel products and tensor products of Markov kernels, but we introduce kernels with special notation that helps with interpreting the resulting objects. String diagrams are the subject of a coherence theorem: taking a string diagram and applying a planar deformation or any of a number of graphical rules not used here yields a string diagram that represents the same kernel (Selinger, 2010). A kernel $\mathbf{A} : X \to \Delta(\mathcal{Y})$ is written as a box with input and output wires, probability measures $\mu \in \Delta(\mathcal{X})$ are written as triangles "closed on the left" and measurable functions $\mathsf{T} : Y \to T$ as triangles "closed on the right". For a thorough definition of version of string diagrams used here, see Cho and Jacobs (2019).



(1)

The identity $\mathbf{Id} : X \to \Delta(X)$ is the Markov kernel $x \mapsto \delta_x$, which we represent with a bare wire. The copy map $\curlyvee : X \to \Delta(X \times X)$ is the Markov kernel $x \mapsto \delta_{(x,x)}$. For $\mathbf{A} : X \to \Delta(Y)$ and $\mathbf{B} : X \to \Delta(Z)$, $\curlyvee(A \otimes B)_x = A_x \otimes B_x$. The discard map $*$ is the Markov kernel $X \to \{\#\}$ given by $x \mapsto \delta_\#$, where # is some one element set.

Given $\mu \in \Delta(X), \mathbf{A} : X \to \Delta(Y)$ as before, the joint distribution on $X \times Y$ that might be informally written $P(\mathsf{X})P(\mathsf{Y}|\mathsf{X})$ is given in string diagram notation on the left of 2. Marginalisation is accomplished with the discard map $*$; hence $\mu\curlyvee(\mathbf{Id} \otimes \mathbf{A}*) = \mu$; this is shown on the right of 2



(2)

3 Statistical Decision Problems and Causal Statistical Decision Problems

A statistical decision problem (SDP) poses the following scenario: suppose we have a set of "states of nature" Θ, a set of decisions D and a loss function $l : \Theta \times D \to \mathbb{R}$. For each state of nature $\theta \in \Theta$ there is an associated probability measure $\mu_\theta \in \Delta(\mathcal{E})$ where $(E, \mathcal{E})$ is some measurable space. Call the stochastic map $\mathbf{H} : \theta \mapsto \mu_\theta$ a *statistical experiment.* Given a *decision strategy* $\mathbf{J} : E \to \Delta(\mathcal{D})$, define the *risk* of $\mathbf{J}$ given state θ to be the expected loss of $\mathbf{J}$ in state θ. Specifically, $R : \Delta(\mathcal{D})^E \times \Theta \to \mathbb{R}$ given by $R : (\mathbf{J}, \theta) \mapsto \mathbf{H}_\theta \mathrm{J} l_\theta$, where we make use of the product notation for brevity.

We would ideally find a strategy $\mathbf{J}$ that minimises the risk in the "true state" θ^*. Unfortunately, we don't know the true state. If there were a decision strategy that minimised the loss in every state, such a strategy would clearly minimise the loss in the true state, but most statistical decision problems don't admit such a strategy. Two alternative decision rules are available:

Given a measure $\xi \in \Delta(\Theta)$ called a prior, ξ-*Bayes decision rule* is a decision rule $\mathbf{J}^*_{\mathrm{Ba}}$ such that the *Bayes risk* $R_\xi : \mathbf{J} \mapsto \int_\Theta H_\theta \mathbf{J} l_\theta d\xi$ is minimised. A *minimax* decision rule $\mathbf{J}^*_{\mathrm{MM}}$ minimises the worst-case risk: $\mathbf{J}^*_{\mathrm{Mm}} \in \arg\min_{\mathbf{J}} \max_{\theta \in \Theta} R(\mathbf{J}, \theta)$ Unlike a Bayes rule, a minimax rule does not invoke a prior. In general, a decision rule is some rule that selects a decision on the basis of the risk functional $R(\mathbf{J}, \cdot)$.

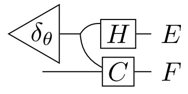
Our representation of statistical experiment is slightly different to, for example, Le Cam (1996), who introduces statistical experiments as an ordered collection of probability measures. Both representations do the same job, and the representation as a map makes for a clearer connection with causal statistical decision problems.

Formally, we define an SDP as the tuple $\langle \Theta, E, D, \mathbf{H}, l \rangle$ where Θ, E and D are measurable sets, $\mathbf{H}$ is a stochastic map $\Theta \to \Delta(\mathcal{E})$ and l a measurable function $E \to \mathbb{R}$. We leave implicit the set of decision strategies $E \to \Delta(\mathcal{D})$. This is a very bare bones exposition of the theory of SDPs, and for more details we refer readers to Ferguson (1967).

Observe that a statistical decision problem supplies a loss l that tells us immediately how desirable a pair $(\theta, d) \in \Theta \times D$ is. It is more typical to talk about how desirable the *consequences* of a decision are than how desirable a (state, decision) pair is. If the set of possible consequences of a decision is denoted by a set F, let the desirability of an element $f \in F$ be given by a utility function $u : F \to \mathbb{R}$. Given such a u, the tuple $\langle \Theta, E, D, \mathbf{H}, u \rangle$ is an ill-posed problem: we want to evaluate the desirability of decision strategies $\mathbf{J}$, but we have no means of connecting decisions with consequences F. We introduce for each state of nature θ a *consequence map* $\mathbf{C}_\theta : D \to \Delta(\mathcal{F})$; let $\mathbf{C}$ be the Markov kernel $\theta \mapsto \kappa_\theta$. We can then define the *causal risk* $S : \Delta(\mathcal{D})^E \times \Theta \to \mathbb{R}$ by $S : (\mathbf{J}, \theta) \mapsto -H_\theta \mathbf{F} C_\theta u$, and Bayes and minimax risks are defined analogously.

For each state $\theta \in \Theta$, the Markov kernel

$$\mathbf{T}_\theta := \qquad (3)$$



Is sufficient to compute the causal risk. Thus we can replace $\mathbf{H}$ and $\mathbf{C}$ with the *causal theory* $\mathbf{T} : \Theta \times D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ given by $(\theta, d) \mapsto \mathbf{T}_\theta(d; \cdot)$. A causal statistical decision problem (CSDP) is therefore a tuple $\langle \Theta, E, F, D, \mathbf{T}, u \rangle$.

Given a CSDP $\alpha = \langle \Theta, E, F, D, T, u \rangle$ where $\mathbf{T}$ is a theory arising from some $\mathbf{H}$ and $\mathbf{C}$ as in Equation 3, we can recover the original kernels by marginalisation: $\mathbf{H} = T(mathbfId \otimes *)$ and $\mathbf{C} = \mathbf{T}(* \otimes \mathbf{Id})$. Given α and letting $l := \mathbf{C}u$ we induce the canonical SDP $\beta = \langle \Theta, E, D, \mathbf{H}', l \rangle$ such that for any $\theta \in \Theta$, $\mathbf{J}$, $R^{(\beta)}(\mathbf{J}, \theta) = S^{(\alpha)}(\mathbf{J}, \theta)$, and thus α and β will always produce identical recommendations.

It is also possible to induce a CSDP from an arbitrary SDP $\beta := \langle \Theta, E, D, \mathbf{H}, l \rangle$. First, define $F := \Theta \times D$ and then let $u := -l$. Define $\mathbf{C} : \Theta \to (D \to \Delta(\mathcal{F}))$ by $\mathbf{C} : \theta \mapsto (d \mapsto (\theta, d))$, and then construct $\mathbf{T}$ from $\Theta, \mathbf{H}$ and $\mathbf{C}$ as in 3. Then the CSDP $\alpha := \langle \Theta, E, F, D, \mathbf{T}, u \rangle$ has the property $S^{(\alpha)}(\mathbf{J}, \theta) = R^{(\beta)}(\mathbf{J}, \theta)$.

Causal theories are the central object of study here. They provide a bridge between the experiment $\mathbf{H}$ and the consequences $\mathbf{C}$ and allow us to use the former to make inferences about the latter.

4 Causal Bayesian Networks

We formulate causal Bayesian networks following Pearl (2009), relying on the source to define the key concept of *compatibility* as it is not a central concern here. We argue for the following claim: a causal Bayesian network represents a causal theory.

Suppose we have a set of "interventions" R which factorises as $R = \times_{i\in[n]}\{\#\} \cup X^i$ for some $n \in \mathbb{N}$, collection of sets $\{X^i\}_{i\in[n]}$ and distinguished element $\# \notin R^i$ for any i. Suppose we also have a measurable space E and set of random variables $\{\mathsf{X}^i | i \in \mathbb{N}\}$ such that $\mathsf{X}^i : E \to X^i$. To explain this setup, an element $(x^0, \#, ..., \#, \#) \in R$ identifies a do intervention where the only non-# component x^0 is the "active" element and the components taking the value # are "do-nothing" elements. Thus $(x^0, \#, ..., \#, \#)$ corresponds to $do(\mathsf{X}^0 = x^0)$ where occurrences of the passive elements are ommitted. Denote by $\underline{\#}$ the element of R consisting entirely of # (equivalently, $do()$).

For $n \in \mathbb{N}$, a directed acyclic graph (DAG) of degree n is a graph $\mathcal{G} = (V, A)$ where V is a set of vertices such that $|V| = n$ and $A \subset V \times V$ is a set of directed edges ("arrows") such that A induces no cycles (for a definition of cycles see Pearl (2009)).

Strictly, we are considering labeled graphs $\mathcal{G}$ and sets $\{\mathsf{X}^i\}_{[n]}$ of random variables. That is, we suppose there is a bijective correspondence between graph nodes and random variables in $\{\mathsf{X}^i\}_{[n]}$ and, leaning on this correspondence, we simply label nodes of $\mathcal{G}$ with the random variables.

We also suppose we have surjective $h : R \to \mathcal{P}([n])$ such that $h : (x^0, ..., x^n) \mapsto \{i | x^i \neq \#\}$. That is, h picks out the active components of r. Define $\mathsf{X}^{i\prime} : R \to \{\#\} \cup X^i$ by the function returning the i-th element of r for $r \in R$. We identify primed and unprimed random variables in the obvious way.

Let $\Pi_{\mathsf{X}^i} := \delta_{\pi_{\mathsf{X}^i}}$ be the Markov kernel that takes the marginal distribution of X^i, and denote by $\mathbf{P}_{r|A}\Pi_{\mathsf{X}^i} : E \to \Delta(\mathcal{X}^i)$ the conditional probability of X^i given the set of random variables A.

Definition 4.1 (Causal Bayesian Network)**.** Given discrete R, E, $\mathbf{P} : R \to \Delta(\mathcal{E})$ and $\{\mathsf{X}^i\}_{i\in[n]}$, a Causal Bayesian Network (CBN) compatible with $\mathbf{P}$ is a directed acyclic graph (DAG) $\mathcal{G}$ of degree n such that for all $r \in R$

1. $\mathbf{P}_r$ is compatible with $\mathcal{G}$ (see Pearl (2009))
2. For all $i \in h(r)$, $\mathbf{P}_r \Pi_{\mathsf{X}^i} = \delta_{\mathsf{X}^{i\prime}(r)}$
3. For all $i \notin h(r)$, $\mathbf{P}_{r|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)}\Pi_{\mathsf{X}^i} = \mathbf{P}_{\underline{\#}|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)}\Pi_{\mathsf{X}^i}$, $\mathbf{P}_{\underline{\#}}$-almost surely

This definition differs slightly from Pearl (2009) in that $\mathbf{P}$ is a Markov kernel rather than a set of labeled elements of $\Delta(\mathcal{E})$.

Given a graph $\mathcal{G}$ and a measure $\mu \in \Delta(\mathcal{E})$ compatible with $\mathcal{G}$ we can define a class of stochastic maps $\mathcal{K} \subset \Delta(\mathcal{E})^V$ such that every $\mathbf{P} \in \mathcal{K}$ is compatible with $\mathcal{G}$ and $\mathbf{P}(\#) = \mu$. Let the notation $\mathcal{G}(\mu)$ stand for the set $\mathcal{K}$ as defined here; note that $\mathcal{G}(\mu)$ is in general set-valued.

We have from this definition for any $r \in V$ the *truncated factorisation* property:

$$P_r F_{\mathbf{X}}(A) = \prod_{i\in h(r)} \delta_{\mathsf{X}^{i\prime}(r)}(\mathsf{X}^i(A)) \sum_{a\in A} \prod_{i\notin h(r)} \mathbf{P}_{\#|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)}\Pi_{\mathsf{X}^i}(a; \{\mathsf{X}^i(a)\}) \qquad (4)$$

As a consequence of the existence of conditional probability for standard measurable spaces, provided μ is compatible with $\mathcal{G}$ we have that the right hand side of4 exists, and so $\mathcal{G}(\mu)$ is non-empty. If μ is positive definite this relationship is functional; in such a case we could treat $\mathcal{G}(\mu)$ as a function from $\Delta(\mathcal{E})$ to interventional maps $\mathbf{P}$.

Suppose we define some arbitrary hypothesis class $\mathcal{H}^{\mathcal{G}} \subset \Delta(\mathcal{E})$ of possible observed distributions. Then let $\Theta := \{(\mu, \mathbf{P}) | \mu \in \mathcal{H}, \mathbf{P} \in \mathcal{G}(\mu)\}$ and define $\mathbf{T}^{\mathcal{G}} : \Theta \times R \to \Delta(\mathcal{E}^2)$ by $(\mu, \mathbf{P}, r) \mapsto \mu \otimes \mathbf{P}_r$. It is natural to consider $\mathbf{T}^{\mathcal{G}}$ the causal theory represented by $\mathcal{G}$ for two reasons: first, it is a natural construction from the definition of a CBN if we take the set of possible *do*-interventions to be the set of decisions for $\mathbf{T}$. Secondly, if we take $\mathcal{H} = \Delta(\mathcal{E})$ then the map from DAGs to causal theories is injective (this is in contrast to, for example, the map from DAGs to probability distributions as in ordinary Bayesian networks(Bishop, 2006)).

Theorem 4.2 (The map $\mathcal{G} \mapsto \mathbf{T}^{\mathcal{G}}$ is injective)**.** *For DAGs $\mathcal{G}$, $\mathcal{G}'$ on the same set of RV's $\{\mathsf{X}^i\}_{[n]}$, $\mathcal{G} \neq \mathcal{G}' \implies \mathcal{T}^{\mathcal{G}} \neq \mathcal{T}^{\mathcal{G}'}$ if these theories are induced by a complete hypothesis class.*

Proof. $\mathcal{G}$ and $\mathcal{G}'$ must disagree on at least one parental set. Suppose this is on the parents of X^i. Choose some μ such that $\mu_{|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)}\Pi_{\mathsf{X}^i} \neq \mu_{|\mathrm{Pa}_{\mathcal{G}'}(\mathsf{X}^i)}\Pi_{\mathsf{X}^i}$. By the non-equality of these conditional probabilities there are some r, r' such that $h(r) = h(r') = \mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i) \cup \mathrm{Pa}_{\mathcal{G}'}(\mathsf{X}^i)$, $\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}'^i)(r) = \mathrm{Pa}_{\mathcal{G}}(\mathsf{X}'^i)(r')$ but $\mathbf{P}^{\mathcal{G}'}_r \Pi_{\mathsf{X}^i} \neq \mathbf{P}^{\mathcal{G}'}_{r'} \Pi_{\mathsf{X}^i}$, but we also have $\mathbf{P}^{\mathcal{G}}_r \Pi_{\mathsf{X}^i} = \mathbf{P}^{\mathcal{G}}_{r'} \Pi_{\mathsf{X}^i}$ by equality of r and r' on $\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)$. Thus $\mathbf{T}^{\mathcal{G}} \neq \mathbf{T}^{\mathcal{G}}$. □

4.1 Decisions and Interventions in a Causal Bayesian Network

The causal theory $\mathbf{T}^{\mathcal{G}}$ associated with a CBN $\mathcal{G}$ features a large number of decisions. Given some utility function $u : \times_{i \in [n]} X^i \to \mathbb{R}$, there is always a decision that fixes the values of all X^i deterministically to a value maximising u, if such a maximum exists. Clearly, for a practical decision problem $\mathbf{T}^{\mathcal{G}}$ is inappropriate. There may be some cases

5 Potential outcomes models

We follow Rubin (2005) for the definition of a potential outcomes model, noting any points of divergence.

We will eschew any discussion of sequences. Following the convention set out in the introduction, we will interchangeably use sans serif letters to refer to "random variables" and "particular strings in the string diagram". W is the treatment assignment taking values in $\{0, 1\}$, $\mathsf{Y}(0)$ $\mathsf{Y}(1)$ are the potential outcomes taking values in Y, Y is the obseved outcome also taking values in Y and X is a "vector of background facts" taking values in X.

Given an underlying state space Θ, a potential outcomes model $\mathcal{O}$ consists of a set of Markov kernels $\langle \mathbf{P}, \mathbf{W}, \mathbf{Y} \rangle$ and a canonical composition that yields a statistical experiment $\mathbf{H} : \Theta \to \Delta(\mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{W})$.

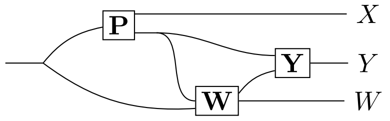
The kernels are

- A "model on the science", $\mathbf{P} : \Theta \to \Delta(\mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{Y})$ (In Rubin's notation, $\mathbf{P}$ is $\prod_i f(\mathsf{X}_i, \mathsf{Y}_i(0), \mathsf{Y}_i(1)|\theta)$)
- An "assignment mechanism", $\mathbf{W} : \Theta \times X \times Y^2 \to \Delta(\{0, 1\})$ (in Rubin's notation, $\mathbf{W}$ is $Pr(\mathsf{W}|X, Y(1), Y(0))$)
- An "observation model", $\mathbf{Y} : \{0, 1\} \times Y^2 \to \Delta(\mathcal{Y})$, defined explicitly as $\mathbf{Y} : (\mathbf{y}^0, \mathbf{y}^1, \mathbf{w}) \mapsto (1 - \mathbf{w}) \odot \delta_{\mathbf{y}^0} + \mathbf{w} \odot \delta_{\mathbf{y}^0}$ where $\odot$ is the elementwise product

We differ from Rubin by defining $\mathbf{Y}$ as a Markov kernel rather than a function. This approach means that we can at best assert $W = w \implies \mathsf{Y} = \mathsf{Y}(w)$ *almost surely* with respect to some probability measure, as a Markov kernel cannot guarantee exact equality. We also differ from Rubin by including Θ in the domain of $\mathbf{W}$ as in our framework leaving this dependence out is equivalent to assuming that the treatment assignment mechanism is known *a priori* (as a result, our state Θ is larger than Rubin's).

We then define the *canonical experiment* $\mathbf{H}^{\mathcal{O}}$ by

$\mathbf{H}^{\mathcal{O}} :=$



(5)

The internal wire from $\mathbf{P}$ to $\mathbf{Y}$ and $\mathbf{W}$ carries the bundle of potential outcomes $\mathsf{Y}(0)\underline{\otimes}\mathsf{Y}(1)$. Apart from the connection between Θ and $\mathbf{W}$, this is consistent with Rubin, though the notation is substantially different.

5.1 Are potential outcomes models causal theories?

By assumption, a potential outcomes model induces a canonical statistical experiment. Given that potential outcomes models are a type of causal model, we can ask whether they induce a canonical *causal theory.* We propose, tentatively, that the answer may be *yes*, though any such construction requires a number of arbitrary decisions (see, for example, **??** for one limitation on the degree to which causal assumptions can be “neutral”). Concretely, we propose that a potential outcomes model corresponds to a causal theory where we are able to decide between every possible treatment map $\mathbf{W}$.

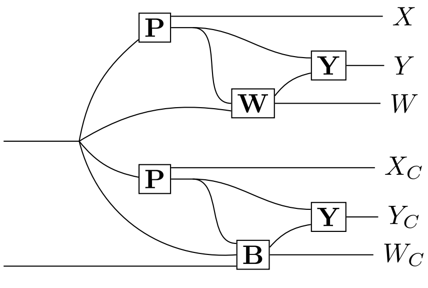
Suppose we have a potential outcomes model $\mathcal{O} := \langle \mathbf{P}, \mathbf{W}, \mathbf{Y} \rangle$ and some rule that maps to a canonical theory $r : \mathcal{O} \mapsto \mathbf{T}^{\mathcal{O}}$. We consider it absolutely necessary that $\mathbf{T}^{\mathcal{O}}(\mathbf{Id} \otimes *) = \mathbf{H}^{\mathcal{O}}$ - that is, both causal theory and PO model agree on the relationship between states and observations. We will proceed on the assumption that the potential outcomes representation of the science $\mathbf{P}$ is consequential. Concretely, we propose that it is desirable that r distinguish any two PO models $\mathcal{O}$ and $\mathcal{O}'$ if these models differ on “the science” $\mathbf{P} \neq \mathbf{P}'$. Any rule not meeting this criterion is asserting that some differences in the science are always unimportant, or equivalently that $\mathcal{P}$ is necessarily an overspecification of the important facts of the problem.

If there is any rule r that meets either the two criteria above, we can induce many more rules that also meet these criteria. For example, take any invertible kernel $\mathbf{A} : F \to \Delta(\mathcal{F})$; then the rule $r' : \mathcal{O} \mapsto r(\mathcal{O})(\mathbf{Id} \otimes \mathbf{A})$ also meets the criteria. Thus we still have arbitrary choices to make given these criteria, and the rule we propose below is additionally motivated by observations of the typical use of potential outcomes.

Suppose $\mathcal{O}$ is equipped with associated discrete spaces $\Theta, X, Y, \{0,1\}$, define $D := [0,1]^{|\Theta|+2|Y|+|X|}$. There is a kernel $\mathbf{B} : D \times \Theta \times X \times Y^2 \to \Delta(\mathcal{W})$ such that for every Markov kernel $\mathbf{W}' : \Theta \times X \times Y^m \to \Delta(\{0,1\})$ there exists $d \in D$ such that $\mathbf{W}' = \mathbf{B}_d$; that is, D indexes the set possible treatment assignment maps. The assumption of discrete spaces is to guarantee the existence of such a $\mathbf{B}$.

From $\mathcal{O}$ and $\mathbf{B}$ we define the *canonical theory* $\mathbf{T}^{\mathcal{O}}$:

$\mathbf{T}^{\mathcal{O}} :=$



(6)

$\mathbf{T}^{\mathcal{O}}$ is two parallel copies of $\mathcal{H}^{\mathcal{O}}$ where $\mathbf{W}$ is replaced by $\mathbf{B}$ in the lower version.

Theorem 5.1. *Given potential outcomes models* $\mathcal{O} = \langle \mathbf{P}, \mathbf{W}, \mathbf{Y} \rangle$, $\mathcal{O}' = \langle \mathbf{P}', \mathbf{W}', \mathbf{Y} \rangle$ *sharing spaces* $\Theta, X, Y, [m]$, *then* $\mathbf{T}^{\mathcal{O}} = \mathbf{T}^{\mathcal{O}'}$ *if and only if* $\mathbf{P} = \mathbf{P}'$ *and* $\mathbf{H} = \mathbf{H}'$.

Proof. Let $\mathbf{T} := \mathbf{T}^{\mathcal{O}}$ and $\mathbf{T}' := \mathbf{T}^{\mathcal{O}'}$.

If $\mathbf{P} = \mathbf{P}'$ and $\mathbf{H} = \mathbf{H}'$ we clearly have $\mathbf{C} := \mathbf{T}(* \otimes \mathrm{Id}) = \mathbf{T}(* \otimes \mathrm{Id})$ as all kernels in the bottom half of 6 ($\mathbf{P}, \mathbf{B}$ and $\mathbf{Y}$) are the same by definition. But then $\mathbf{T} = \curlyvee(\mathbf{H} \otimes \mathbf{C}) = \mathbf{T}'$.

Suppose $\mathbf{T} = \mathbf{T}'$ and $\mathbf{P} \neq \mathbf{P}'$. Then there exists some $A \in \mathcal{X} \otimes \mathcal{Y}^2$, $\theta \in \Theta$ such that $\mathbf{P}_\theta(A) \neq \mathbf{P}'_\theta(A)$. Choose $d \in D$ such that $\mathbf{B}_{\theta,d,x,y_0,y_1} = \delta_0$ if $(x, y_0, y_1) \in A$ and $\mathbf{B}_{\theta,d,x,y_0,y_1} = \delta_1$ otherwise. But then $\mathbf{T}^{\mathcal{O}}_{\theta,d}\pi_{\mathsf{W}}(\{0\}) = \mathbf{P}_\theta(A) \neq \mathbf{P}'_\theta(A) = \mathbf{T}^{\mathcal{O}'}_{\theta,d}\pi_{\mathsf{W}}(\{0\})$, a contradiction. Thus $\mathbf{P} = \mathbf{P}'$. In addition, $\mathbf{H} = \mathbf{T}(\mathrm{Id} \otimes *) = \mathbf{H}'$. □

Note that the assignment $\mathbf{W}$ may differ between $\mathcal{O}$ and $\mathcal{O}'$. Suppose $X = \emptyset$, $Y = \{0,1\}$ and for some θ, $\mathbf{P}_\theta = \frac{1}{4}(\delta_{0,0} + \delta_{0,1} + \delta_{1,0} + \delta_{1,1})$. Then $\mathbf{W}_\theta : (y_0, y_1) \mapsto [\![y_0 = y_1]\!]\delta_0 + [\![y_0 \neq y_1]\!]\delta_1$ and $\mathbf{W}'_\theta := 1 - \mathbf{W}_\theta$ both yield the same observations $\mathbf{H}_\theta$. It may be desirable that a mapping from PO models to causal theories also distinguishes PO models that differ only on $\mathbf{W}$ - for example, we might want some decision $d \in D$ to always be interpretable as “in the state θ, raise the probability of treatment above the observed level iff $y_0 \neq y_1$”, a decision which would yield different consequences given a theory based on $\mathbf{W}_\theta$ or $\mathbf{W}'_\theta$.

6 Dextrous Theories and Pragmatic Theories

While we began by defining causal theories as objects to be used in the solution of CSDPs, we have for both potential outcomes models and CBNs constructed theories that are clearly unfit for this task. Under $\mathbf{T}^{\mathcal{O}}$ we have the possibility (among others) of deciding to assign treatment if and only if $y_1 > y_0$. Under a CBN theory, we can simply apply a $do()$ intervention that sets all variables to their most desirable state; in neither case is any inference warranted whatsoever. We will call theories that posit unreasonably fine levels of control *dextrous theories.*

For the purpose of choosing a strategy to solve a decision problem, we want to work with a *pragmatic theory* describing a more restricted set of decisions that correspond to actions we believe we can actually take. However, we may be willing to accept that a dextrous theory $\mathbf{T}^x$ models our pragmatic theory $\mathbf{T}^g$ in the sense that there is a correspondence m between decisions in each theory such that the consequences of d via $\mathbf{T}^g$ are the same as the consequences of $m(d)$ via $\mathbf{T}^x$. In this sense a dextrous theory may still be useful for strategy choice in real decision problems. To illustrate this, consider the problem of evaluating the “effect of assigning treatment” vs “the effect of receiving treatment”. From Shrier et al. (2017):

> In public health, we are normally concerned with the first question – the effect of assigning a treatment. If we implement a prevention or treatment program that is efficacious only under strict research conditions but people in the real world would not receive it for any possible reason, the program will not be effective. This real-world context is termed the "average causal effect" of assigning treatment and is best estimated by the intention-to-treat (ITT) analysis [...]
>
> There are 2 reasons why the average causal effect ofreceiving a treatment may be more important than the ITT for some people. First, even in the public health domain, investigators may want to know what the average causal effect of a treatment program would be if they could improve participation in the program. [...] Also, the average causal effect of receiving a treatment is of primary interest to a patient deciding whether or not to take the treatment as recommended.

Concretely, suppose we have two dextrous theories $\mathbf{T}^{\text{ITT}}$ and $\mathbf{T}^{\text{RT}}$ modelling effects of intention-to-treat and receiving treatment respectively, and we want pragmatic theories $\mathbf{T}^p, \mathbf{T}^t$ describing the effects of prescribing treatment and taking treatment respectively. Consider the first decision problem described: we have two pragmatic decisions $D^p = \{0,1\}$ where $d_p = 1$ corresponds to "implement a treatment program" and $d_p = 0$ corresponds to "do nothing". Under the intention to treat model $\mathbf{T}^{\text{ITT}}$ we are willing to accept that these decisions correspond determinisically setting $\mathsf{W} = 1$ and $\mathsf{W} = 0$ respectively. That is, we suppose that the consequence of choosing $d_p = 1$ in the pragmatic theory $\mathbf{T}^p$ is the same as the consequence of choosing the decision $e_1 \in D^{\text{ITT}}$ such that for all θ, x, y_0, y_1, $\mathbf{B}^{\text{ITT}}_{\theta,d,x,y_0,y_1} = \delta_1$ and analogously $d_p = 0$ corresponds to the elemet $e_0 \in D^{\text{ITT}}$ that sets W to 0.

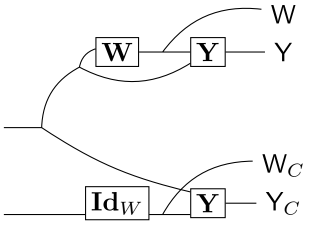
Consider the same problem – that of implementing a treatment program – for the dextrous theory of receiving treatment $\mathbf{T}^{\text{RT}}$. Here, a correspondence between D^p and D^{RT} is less clear. We may accept that implementing a treatment program corresponds to *some* choice of treatment taking function $\mathbf{B}^{\text{RT}}$, one that is perhaps more likely to result in treatment than that for doing nothing. However, without more information we have only the idea that there is a correspondence between D^{RT} and D^p and not what this correspondence is. We could, of course, express our uncertainty with a set of possible correspondences or a probability measure over correspondences and proceed from there.

Consider the third decision problem, where decisions correspond to taking the treatment, and in particular consider using the dextrous theory $\mathbf{T}^{\text{ITT}}$. It is very likely that *no* choice of prescription function W^{ITT} is consistent with the test subjects always taking the treatment. That is, we're not just uncertain about the correspondence between dextrous decisions D^{ITT} and pragmatic decisions D^t - we are in fact confident that there is no such correspondence.

7 Comparing Causal Bayesian Networks and Potential Outcomes theories

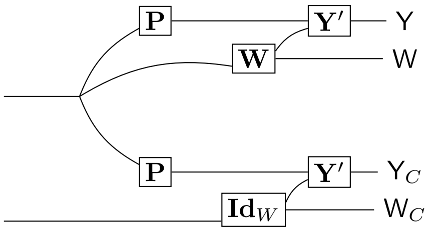
Expressing both CBNs and PO models as causal theories allows us to compare models from framework, and we see that causal theories typically obtained are rather different. For example, a CBN $\mathcal{G}$ defines an intervention operation for every random variable that has been represented as a node of $\mathcal{G}$ while a PO model $\mathcal{O}$ (at least in the version developed here) will typically not allow any decisions that deterministically set Y and no decisions may affect X at all. On the other hand, decisions in a theory $\mathbf{T}^{\mathcal{O}}$ may yield arbitrary dependence of W on a number of unobserved quantities, which is not a possibility at least for the basic type of CBN discussed here. However, it may be the case that the pragmatic theory we derive starting from either a CBN or PO theory is the same.

Suppose we have a CBN $\mathcal{G} := W \rightarrowtriangle Y$, where W and Y are random variables taking values in some arbitrary spaces W and Y. Suppose also that we require a pragmatic theory $\mathbf{T}^{\mathcal{G}} : \Theta \times W \to \Delta([\mathcal{W} \otimes \mathcal{Y}]^2)$ where our decisions correspond only to "hard do interventions" $do(W = w)$ on W under the full CBN theory – that is, we have no decisions corresponding to do-interventions on Y or do-nothing. Then there exist Markov kernels $\mathbf{W} : \Theta \to \Delta(\mathcal{W})$, $\mathbf{Y} : \Theta \times W \to \Delta(\mathcal{Y})$ such that $\mathbf{T}^{\mathcal{G}}$ can be represented as in the diagram 7. Conversely, any causal theory that can be represented in this manner is a candidate for $\mathbf{T}^{\mathcal{G}}$ (see A)



(7)

Suppose we have a PO model $\mathcal{O} = \langle \mathbf{P}, \mathbf{W}, \mathbf{Y} \rangle$ on Θ, W and Y such that $\mathbf{W}$ depends only on Θ. Suppose also that we require a pragmatic theory where, similarly to the case above, decisions correspond only to "setting" W in the standard theory associated with $\mathcal{O}$. Then the resulting theory can be represented by the diagram 8. Conversely, there is a PO model for every causal theory with this representation.



(8)

In the lower diagram we can define $\mathbf{Y}^* := (\mathbf{P} \otimes \mathbf{Id})\mathbf{Y}'$ to produce a diagram that is topologically equivalent to 7. While $\mathbf{P}$ and $\mathbf{W}$ are arbitrary, $\mathbf{Y}'$ has a particular form. It is still possible to express a general kernel $\mathbf{Y} : \Theta \times W \to \Delta(\mathcal{Y})$ in the form of $\mathbf{Y}^*$; let $\mathbf{P} : \theta \mapsto \mathbf{Y}_{\theta,0} \otimes \mathbf{Y}_{\theta,1}$. Then $\mathbf{Y}^* = \mathbf{Y}$. Thus under the restrictions given, the sets of viable pragmatic theories derived from the CBN and the PO model are exactly the same.

7.1 Coarsening and Saved Inference

The causal theories associated with both CBN and PO models are very profligate. They define many decisions that are unlikely to be considered in a pragmatic decision problem, and in practice it is usually only possible to determine the consequences of a small subset of these decisions if it is possible to determine any at all. In additon, proponents of both theories have advocated for the universality of the "causal effects" they represent:

> The perspective that (1) the science exists independently of how we try to learn about it and that (2) if the model used for analysis of the resulting data is approximately correct, then the resulting posterior distribution will give a fair summary of the current state of knowledge of that science seems, at least to me, consistent with common views of the scientific enterprise [...] The potential outcomes, together with covariates, define the science in the sense that all causal estimands are functions of these values (Rubin, 2005)

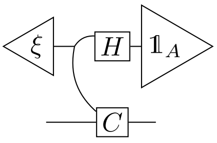
> By representing the domain in the form of an assembly of stable mechanisms, we have in fact created an oracle capable of answering queries about the effects of a huge set of actions and action combinations (Pearl, 2009)

We present here a somewhat speculative account of what both of these approaches are trying to achieve based on the notions of *coarsening* and *saved inference*. The basic story is: suppose our job is to study the causal dynamics of some system, but we're not quite sure of who will put our results to use or what decisions they will have available. We adopt a theory $\mathbf{T}$ with the hope that we can pass along results to end users, even if they are operating with a more pragmatic theory $\mathbf{T}'$. In this section we specify precisely when it is possible to do so.

Definition 7.1 (Coarsening)**.** A theory $\mathbf{T} : \Theta \times D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ can be coarsened to a theory $\mathbf{T}' : \Theta \times D' \to \Delta(\mathcal{E} \otimes \mathcal{F})$ if there exists $M : D' \to \Delta(\mathcal{D})$ such that $\mathbf{T}' = (\mathbf{Id} \otimes M)\mathbf{T}$. We say that $\mathbf{T}'$ is *clumsier* than $\mathbf{T}$ or $\mathbf{T}$ is *more dextrous* than $\mathbf{T}'$.

Given $\mathbf{T}$, an event $A \in \mathcal{E}$ with $\xi H \mathbb{1}_A > 0$, write the theory conditioned on A as $\mathbf{T}_\xi|A : D \to \Delta(\mathcal{F})$, defined as

$$\mathbf{T}_\xi|A := (\xi H(A))^{-1}$$



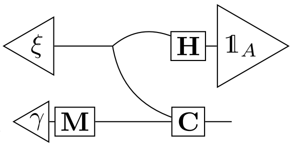
(9)

Note that $\mathbf{T}_\xi|A$ along with a strategy $\gamma \in \Delta(\mathcal{D})$ is the conditional probability of F by the elementary definition - for $B \in \mathcal{F}$, $\mathbf{T}_{\xi,\gamma}|A : B \mapsto \frac{(\xi \otimes \gamma)\mathbf{T}(A,B))}{\xi \mathbf{H}(A)}$.

Theorem 7.2. *Given* $\mathbf{T} : \Theta \times D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ *and* $\mathbf{T}' : \Theta \times D' \to \Delta(\mathcal{E} \otimes \mathcal{F})$*, there exists* $\mathbf{M}$ *such that* $\mathbf{T}'_{\gamma,\xi}|A = \gamma M \mathbf{T}_\xi|A$ *for all* $\xi \in \Delta(\Theta)$*,* $\gamma \in \Delta(\mathcal{D}')$ *and* $A \in \mathcal{E}$ *where* $\xi \mathbf{H}(A) > 0$ *if and only if* $\mathbf{T}'$ *is a coarsening of* $\mathbf{T}$*.*

Proof. Let the coarsening from $\mathbf{T}$ to $\mathbf{T}'$ be witnessed by $\mathbf{M} : D' \to \Delta(\mathcal{D})$. For arbitrary ξ, A such that $\xi \mathbf{H}(A) > 0$ and arbitrary γ:

$$\gamma \mathbf{M} \mathbf{T}_\xi|A = (\xi \mathbf{H}(A))^{-1}$$



(10)

$$= (\xi \mathbf{H}(A))^{-1} \gamma (\xi \otimes \mathbf{M}) \mathbf{T} (\mathbb{1}_A \otimes \mathrm{Id}_F) \quad (11)$$

$$= (\xi \mathbf{H}(A))^{-1} \gamma (\xi \otimes \mathrm{Id}_D) \mathbf{T}' (\mathbb{1}_A \otimes \mathrm{Id}_F) \quad (12)$$

$$= \mathbf{T}'_{\gamma,\xi}|A \quad (13)$$

Suppose we have $\mathbf{M}$ such that $\mathbf{T}'_{\gamma,\xi}|A = \gamma M \mathbf{T}_\xi|A$ for all $\xi \in \Delta(\Theta)$, $\gamma \in \Delta(\mathcal{D}')$ and $A \in \mathcal{E}$ where $\xi \mathbf{H}(A) > 0$. Choose some arbitrary θ. Then for *all* $A \in \mathcal{E}$ either $\mathbf{H}_\theta(A) = 0$ in which case $\mathbf{M}\mathbf{T}_{\theta,d'}(A \times B) = \mathbf{T}'_{\theta,d'}(A \times B) = 0$ for all $d' \in D', B \in \mathcal{F}$ or $\mathbf{H}_\theta(A) > 0$ in which case for all d', B:

$$(\mathbf{H}_\theta(A))^{-1}(\delta_\theta \otimes \delta_{d'})\mathbf{T}'(\mathbb{1}_A \otimes \mathbb{1}_B) = (\mathbf{H}_\theta(A))^{-1}(\delta_\theta \otimes \delta_{d'}\mathbf{M})\mathbf{T}(\mathbb{1}_A \otimes \mathbb{1}_B) \tag{14}$$

$$(\delta_\theta \otimes \delta_{d'})\mathbf{T}'(\mathbb{1}_A \otimes \mathbb{1}_B) = (\delta_\theta \otimes \delta_{d'}\mathbf{M})\mathbf{T}(\mathbb{1}_A \otimes \mathbb{1}_B) \tag{15}$$

Hence

$$(\mathbf{Id} \otimes \mathbf{M})\mathbf{T} = \mathbf{T}' \tag{16}$$

□

In other words, if and only if $\mathbf{T}$ can be coarsened to $\mathbf{T}'$ then we can "save" the results of conditioning $\mathbf{T}_\xi$ on A via $\mathbf{T}_\xi|A$ and later on we can determine the effects of some strategy γ on $\mathbf{T}'$ via $\gamma\mathbf{M}\mathbf{T}_\xi|A$.

Recall our discussion of the problems of determining the effects of a treatment program vs determining the effects of taking the treatment. We had established that there was known correspondence between D^{ITT} and D^p – this correspondence was, precisely, a coarsening from $\mathbf{T}^{\text{ITT}}$ to $\mathbf{T}^p$ (concretely, it proceeds via the kernel $\mathbf{M} : 1 \mapsto \delta_{e_1}$ and $0 \mapsto \delta_{e_0}$). On the other hand, while we didn't identify a known coarsening from $\mathbf{T}^{\text{RT}}$ to $\mathbf{T}^p$, we were argued that such a coarsening likely existed. Finally, we rejected the possibility of a coarsening from $\mathbf{T}^{\text{ITT}}$ to $\mathbf{T}^t$.

Note that coarsening doesn't require a functional relationship between D' and D – a stochastic map $D' \to \Delta(\mathcal{D})$ is good enough. Thus a coarsening can represent uncertainty over which decisions correspond. This is another potential reason for the profligate decision sets associated with the canonical causal theories we have discussed - it is desirable to feature decisions that we might be certain correspond to decisions in some realistic theory, and also decisions that *might* correspond.

7.2 Universal Coarsening

We postulate on the basis of the quoted statements from Pearl and Rubin as well as the nature of the causal theories associated with CBNs and PO models that these modelling approaches aim to yield "fantastical" causal theories that aren't only useful for solving particular decision problems, but can be coarsened to a wide variety of useful realistic theories. We conclude this discussion with a result showing that there are limits on this enterprise. The only causal theories that can be universally coarsened are consequentially saturated. In this case, we have a theory that offloads all responsibility for informing about consequences onto the choice of coarsening kernel $\mathbf{M}$.

Theorem 7.3 (Universal Coarsening)**.** *Fixing discrete spaces Θ, D^*, E, F, define a* saturated *causal theory $\mathbf{T}^* : \Theta \times D^* \to \Delta(\mathcal{E} \otimes \mathcal{F})$ as a theory where, for every $\mathbf{K} : \Theta \to \Delta(\mathcal{F})$, there exists $\gamma \in \Delta(D^*)$ such that $\mathbf{K} = (\mathbf{Id} \otimes \gamma)\mathbf{T}^*$.*

The following statements are equivalent:

- $\mathbf{T}^*$ *is a saturated theory*
- *For all discrete D and all theories $\mathbf{T} : \Theta \times D \to \Delta(\mathcal{E} \otimes \mathcal{F})$, $\mathbf{T}$ is a coarsening of $\mathbf{T}^*$.*

Proof. Suppose $\mathbf{T}^*$ is saturated. For every $d \in D$ there is some $f(d) \in D^*$ such that $\mathbf{T}_d = \mathbf{T}_{f(d)}$. Let $\mathbf{M} : d \mapsto \delta_{f(d)}$; then $\mathbf{T}$ is a coarsening of $\mathbf{T}^*$ by $\mathbf{M}$.

If every theory is a coarsening of $\mathbf{T}^*$, then in particular the saturated theory $\mathbf{T}^{\text{sat}} : \Theta \times D^* \to \Delta(\mathcal{E} \otimes \mathcal{F})$ is a coarsening of $\mathbf{T}^*$. But then there is some $\mathbf{M}$ such that $(\mathbf{Id} \otimes \mathbf{M})\mathbf{T}^* = \mathbf{T}^{\text{sat}}$, so $\mathbf{T}^*$ must also be saturated. □

8 Discussion

We have introduced an original approach to formulating questions of causal inference and analysing approaches to causal modelling. We take cues from statistical decision theory in the realm of problem definition and make heavy use of the theory of Markov kernels for reasoning about causal theories, the central object of our approach. Our approach makes crystal clear the distinction between "statistical" and "causal" knowledge – the former is represented by a statistical experiment and the latter by a causal theory. We can also plausibly interpret the two major existing approaches of Causal Bayesian Networks and Potenial Outcomes as tools to generate causal theories, though there are arbitrary decisions that must be made in order to do this.

Though we develop this theory in the context of "small world decision problems" (Joyce, 1999), we also make progress on the question of what causal theories are doing apart from facilitating reasoning about small world decision problems. We show that if a potentially unrealistic theory can be related to a more realistic theory by coarsening, then knowledge of consequences under the former may be informative about consequences under the latter.

While we do not address the unique questions that can be raised with counterfactual models (Pearl, 2009), our approach suggests an alternative view for the relationship between counterfactual and interventional causal models. Rather than occupying different levels of a hierarchy, each yields causal theories with different kinds of rich decisions sets. It is plausible that the different sets of decisions each approach provides may be amenable to coarsening in different domains. Indeed, we see extensive discussion of counterfactual treatment effects in the econometrics literature, where decisions usually involve changing incentives which can plausibly be understood as altering the assignment function $\mathbf{W}$ in unpredictable ways (Angrist and Pischke, 2014; Carneiro et al., 2010; Imbens and Angrist, 1994). Causal Bayesian Networks, on the other hand, have found applications in the study of biological systems which typically feature large numbers of variables which permit a wide variety of targetted interventions (Sachs et al., 2005; Maathuis et al., 2009).

While Theorem 7.2 suggests that coarsening can be useful for "reusing knowledge" between compatible causal theories, this is only likely to be helpful if it is possible to determine that a theory $\mathbf{T}'$ is a coarsening of $\mathbf{T}$ under $\mathbf{M}$ without having to perform inference on both $\mathbf{T}$ and $\mathbf{T}'$ to satisfy ourselves that the consequences do indeed match in detail for both theories. Understanding when we can consider $\mathbf{T}'$ to be a coarsening of $\mathbf{T}$ and when it is useful to do so is an important development of the ideas presented here. Informally, we want to understand the question "if I know my decision definitely results in $\mathsf{X} = x$, when do I also know it corresponds to $do(\mathsf{X} = x)$?"

A number of the results here are predicated on discrete spaces, a step that allows us to disregard questions of measurability. A second important direction of development is extending this theory to continuous spaces and understanding what limitations this introduces. Relatedly, the notions of conditional probability, conditioning, independence and Bayesian inversion are well understood in the context of probability measures, including in their string diagrammatic treatment (Cho and Jacobs, 2019), but we are not aware of analogues of these notions for general Markov kernels, if they exist. They would be invaluable tools in the analysis of causal theories, which, owing to the dependence on D, are not naturally dealt with as probability measures.

The string diagram notation we use has a strong connection with the DAGs (Fong, 2013) used in causal graphical models as well as to influence diagrams(Dawid, 2002), as do Markov kernels themselves. It would not be surprising if there were a deep connection between the two.

References

Joshua D. Angrist and Jörn-Steffen Pischke. *Mastering 'Metrics: The Path from Cause to Effect*. Princeton University Press, Princeton ; Oxford, with french flaps edition edition, December 2014. ISBN 978-0-691-15284-4.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant Risk Minimization. *arXiv:1907.02893 [cs, stat]*, July 2019. URL `http://arxiv.org/abs/1907.02893`. arXiv: 1907.02893.

Christopher Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer-Verlag, New York, 2006. ISBN 978-0-387-31073-2. URL `https://www.springer.com/gp/book/9780387310732`.

Pedro Carneiro, James J. Heckman, and Edward Vytlacil. Evaluating Marginal Policy Changes and the Average Effect of Treatment for Individuals at the Margin. *Econometrica*, 78(1):377–394, 2010. ISSN 1468-0262. doi: 10.3982/ECTA7089. URL `http://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA7089`.

Erhan Çinlar. *Probability and Stochastics*. Springer, 2011.

Kenta Cho and Bart Jacobs. Disintegration and Bayesian inversion via string diagrams. *Mathematical Structures in Computer Science*, 29(7): 938–971, August 2019. ISSN 0960-1295, 1469-8072. doi: 10.1017/S0960129518000488. URL `https://www.cambridge.org/core/journals/mathematical-structures-in-computer-science/article/disintegration-and-bayesian-inversion-via-st 0581C747DB5793756FE135C70B3B6D51`.

A. P. Dawid. Influence Diagrams for Causal Modelling and Inference. *International Statistical Review*, 70(2):161–189, 2002. ISSN 1751-5823. doi: 10.1111/j.1751-5823.2002.tb00354.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1751-5823.2002.tb00354.x`.

Thomas S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, July 1967. ISBN 978-1-4832-2123-6.

Brendan Fong. Causal Theories: A Categorical Perspective on Bayesian Networks. *arXiv:1301.6201 [math]*, January 2013. URL `http://arxiv.org/abs/1301.6201`. arXiv: 1301.6201.

Guido W. Imbens and Joshua D. Angrist. Identification and Estimation of Local Average Treatment Effects. *Econometrica*, 62(2):467–475, 1994. ISSN 0012-9682. doi: 10.2307/2951620. URL `https://www.jstor.org/stable/2951620`.

James M. Joyce. *The Foundations of Causal Decision Theory*. Cambridge University Press, Cambridge ; New York, April 1999. ISBN 978-0-521-64164-7.

L. Le Cam. Comparison of Experiments - A Short Review.pdf. *IMS Lecture Notes - Monograph Series*, 30, 1996.

Marloes H. Maathuis, Markus Kalisch, and Peter Bühlmann. Estimating high-dimensional intervention effects from observational data. *The Annals of Statistics*, 37(6A):3133–3164, December 2009. ISSN 0090-5364, 2168-8966. doi: 10.1214/09-AOS685. URL `https://projecteuclid.org/euclid.aos/1250515382`.

William Edward Morris and Charlotte R. Brown. David Hume. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2019 edition, 2019. URL `https://plato.stanford.edu/archives/sum2019/entries/hume/`.

Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition, 2009.

Thomas S Richardson and James M Robins. Single world intervention graphs (swigs): A unification of the counterfactual and graphical approaches to causality. *Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper*, 128(30):2013, 2013.

Donald B. Rubin. Causal Inference Using Potential Outcomes. *Journal of the American Statistical Association*, 100(469):322–331, March 2005. ISSN 0162-1459. doi: 10.1198/016214504000001880. URL `https://doi.org/10.1198/016214504000001880`.

Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. *Science*, 308(5721):523–529, April 2005. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1105809. URL `https://science.sciencemag.org/content/308/5721/523`.

Leonard J. Savage. *Foundations of Statistics*. Dover Publications, New York, revised edition edition, June 1972. ISBN 978-0-486-62349-8.

Peter Selinger. A survey of graphical languages for monoidal categories. *arXiv:0908.3347 [math]*, 813: 289–355, 2010. doi: 10.1007/978-3-642-12821-9_4. URL `http://arxiv.org/abs/0908.3347`. arXiv: 0908.3347.

Ilya Shpitser. *Complete Identification Methods for Causal Inference*. PhD Thesis, University of California at Los Angeles, Los Angeles, CA, USA, 2008.

Ian Shrier, Evert Verhagen, and Steven D. Stovitz. The Intention-to-Treat Analysis Is Not Always the Conservative Approach. *The American Journal of Medicine*, 130(7):867–871, July 2017. ISSN 1555-7162. doi: 10.1016/j.amjmed.2017.03.023.

Abraham Wald. *Statistical decision functions*. Statistical decision functions. Wiley, Oxford, England, 1950.

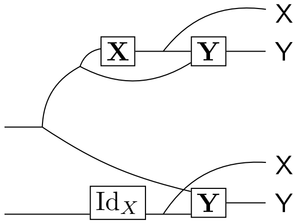
Supplement to: AIStats Submission Sept 26: up to "CBNs are causal theories"

A CBN representation as a causal theory

Definition A.1 (Elementary Causal Bayesian Network)**.** Given D, E, Θ, random variables $\{\mathsf{X}^i\}_{i\in[n]}$ on E, a distinguished variable X^0 taking values in D and a causal theory $T : \Theta \times D \to \Delta(\mathcal{E} \otimes \mathcal{E})$ with $H := T(\mathrm{Id}_E \otimes *_E)$ and $C := T(*_E \otimes \mathrm{Id}_E)$, an *elementary Causal Bayesian Network* (eCBN) compatible with T is a directed acyclic graph (DAG) $\mathcal{G}$ with nodes $\{\mathsf{X}^i\}_{i\in[n]}$ such that

1. H_θ and $C_{\theta,d}$ are compatible with $\mathcal{G}$ (see Pearl (2009))
2. $C_{\theta,d}F_{\mathsf{X}^i} = \delta_d$
3. For all $i \neq 0$, $C_{\theta|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)}F_{\mathsf{X}^i} = H_{\theta|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)}F_{\mathsf{X}^i}$, H_θ-almost surely

Suppose we have the EDAG $\mathcal{G} := X \rightarrowtriangle Y$, where X and Y are random variables taking values in some arbitrary spaces X and Y. Then $\mathcal{G}$ is compatible with a causal theory $T : \Theta \times X \to \Delta([\mathcal{X} \otimes \mathcal{Y}]^2)$ if and only if there exist Markov kernels $\mathbf{X} : \Theta \to \Delta(\mathcal{X})$, $\mathbf{Y} : \Theta \times X \to \Delta(\mathcal{Y})$ such that



(17)

Here we represent the identity kernel explicitly to make clear that it replaces $\mathbf{X}$ in the lower part of the diagram. This fact is hidden by the usual convention of representing the identity by a bare wire.

Proof. Condition 1 is vacuous. Condition 2 is equivalent to asserting that $C_{\theta,d}F_{\mathsf{X}^i}$ is the identity map. The shared kernel $\mathbf{Y}$ guarantees 3 and if $\mathbf{Y}$ cannot be shared then 3 does not hold. □

A particularly interesting feature of this representation is the fact that the edge cutting behaviour, usually an implicit part of the definition of a CBN, is displayed explicitly by replacing $\mathbf{X}$ by the identity.