

---

# AISTats Submission Sept 26: up to "CBNs are causal theories"

---

## 1 Introduction

It is widely accepted that causal knowledge and statistical knowledge are distinct. At least two levels are common: statistics is concerned with *association* while causation is concerned with *consequences*; a distinction of this nature goes back at least to Hume [Morris and Brown, 2019], who is also noted for his argument that knowledge of the latter cannot be reduced to the former. Pearl [2009] has identified three level hierarchy of causal knowledge in contemporary work: first *association*, then *intervention* (analogous to Cartwright's *strategy*) and finally *counterfactuals*. Pearl argues that the types of things that can be known at higher levels subsumes what can be known at lower levels (e.g. all associational knowledge is a type of interventional knowledge), but lower levels do not subsume higher levels.

An apparently paradoxical feature of this three level hierarchy is that, though knowledge is claimed to flow only in one direction, we find that the first and third levels are both described by ordinary joint probability distributions. Counterfactual queries can be formulated as missing data problems, which are distinct from associational problems only due to the interpretations we assign to so-called *counterfactual random variables* or *potential outcomes*. Knowledge at the second level, on the other hand, is described by causal graphical models which are *not* joint probability distributions (in one treatment, they are introduced as indexed sets of joint probability distributions Pearl [2009]). Here is an apparently paradoxical feature of common approaches to causal inference: associational knowledge is distinguished from consequential knowledge in both interpretation and representation, while counterfactual knowledge – considered to subsume both – is distinguished from associational knowledge by interpretation only.

Statistical decision theory, introduced by Wald [1950], underpins much of modern machine learning. It introduced the fundamental notions of *loss* and *risk* to statistics and provided foundational theorems such as the *complete class theorem* which shows that all admissible decision rules are Bayesian decision rules for some prior. Key elements of statistical learning theory inherits heavily from statistical decision theory. While some descendants of statistical decision theory have grappled with the problem of causality [Lewis, 1981], SDT itself is regarded as a theory of statistical decision making and not of causality.

We show a surprising relationship between SDT and causal graphical models. We proceed in two steps: We note that a causal graphical model represents a relationship between probability measures on a given space and the consequences of a given set of actions. We then consider a modification of a standard statistical decision problem: suppose that, rather than being given a loss function that directly evaluates decisions, we are instead provided with a preference function over consequences of decisions that (following convention) we call a *utility*. The resulting problem is underspecified and provides no ordering over decisions. However, the type of relationship represented by a causal graphical model is then found to be precisely the type of object needed to fully specify the problem, and does so in a way that induces a regular statistical decision problem.

This motivates the definition of *causal statistical decision problems* (CSDPs). These relate to regular statistical decision problems (SDPs) in loose analogy with the way that model based reinforcement learning relates to model free reinforcement learning; while the former keeps track of both consequences and rewards/utilities of decisions, the latter “forgets about the consequences” and only works with reward/utility.

Is this true?  
There are substantial similarities between SDT and SLT, but I haven't found direct evidence of lineage in e.g. a citation from Valiant

CSDPs introduce the notion of *causal theories*. Causal theories represent relationships between probability measures and consequences and are a generalisation of causal graphical models. In Pearl’s language, they represent the connection between associational knowledge and interventional knowledge; in Cartwright’s, they connect associational knowledge with the consequences of strategies.

Thanks to the clarity of our approach, we are able to shed light on the questions raised in the second paragraph: we require a causal theory to bring knowledge from levels 1 to 2 of Pearl’s hierarchy and we *also* require a causal theory to bring knowledge from level 3 to level 2. A joint distribution over counterfactuals can only answer interventional questions *given interventional assumptions* (we speculate that such assumptions may have been taken for granted). Associational knowledge is represented with probability distributions, knowledge of consequences with stochastic maps and relationships between the two with causal theories.

Choosing appropriate causal theories is a hard problem. Whether we build a causal theory with graphical models or Potential Outcomes (with additional assumptions), it is often the case that a nontrivial result rests on assumptions that are not obvious, generic or testable. Generic principles such as the bias-variance tradeoff have proved to be immensely powerful in the world of statistics, and we regard the question of whether there are generic principles that govern causal inference and what they may be to be one of the most important questions in the field.

We are primarily concerned with setting out a clear framework for reasoning about causal theories, and do not propose principles for constructing a causal theory in this paper. We are able to show a general negative result - causal theories that are symmetric over permutations of decisions cannot yield nontrivial decision rule orderings. We term this result “no causes in, no causes out” as it demonstrates that some causal knowledge is required at the outset if we hope for any nontrivial decision rules. Such asymmetric causal assumptions must be problem specific, so from the outset we cannot build causal theories on “problem neutral” assumptions alone.

There’s another half baked angle here, which is “what kinds of causal theories are represented by graphical models”? In particular, via the question of dominance we can consider causal theories to be related by three different types of randomisation. Also, if we examine marginal causal models, we note that they all represent causal theories that are related to a “nice” causal theory (in the sense that identification is straightforward) via two of these types of randomisation. It’s half baked because I can’t yet say a lot from there, save for the fact that the operation of randomisation seems more amenable to being generalised to a continuous version than DAGs do.

I could also include the “free” results from statistical decision theory somewhere - complete class theorem, purification

## 2 Definitions & Notation

We use the following standard notation:  $[n]$  refers to the set of natural numbers  $\{1, \dots, n\}$ . Sets are ordinary capital letters  $X$ ,  $\sigma$ -algebras are script letters  $\mathcal{X}$  while random variables are sans serif capitals  $X : \_ \rightarrow X$ . The calligraphic  $\mathcal{G}$  refers to a directed acyclic graph rather than a  $\sigma$ -algebra. Probability measures are greek letters  $\mu, \xi, \gamma$  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  $\Delta(\mathcal{X})$  and the set of all stochastic maps  $W \rightarrow \Delta(X)$  as  $\Delta(\mathcal{X})^W$ .  $\delta_x : (X) \rightarrow [0, 1]$  is the probability measure such that  $\delta_x(A) = 1$  if  $x \in A$  and 0 otherwise.

If  $X$  is a discrete space, probability measures on  $X$  are positive row vectors in  $\mathbb{R}^{|X|}$  that sum to 1, and stochastic maps or Markov kernels  $X \rightarrow \Delta(Y)$  are  $|X| \times |Y|$  positive matrices with row sums of 1. Using the standard notion of associative matrix-vector products, given  $\mu \in \Delta(X)$  and  $\mathbf{A} : X \rightarrow \Delta(Y)$ ,  $\mu\mathbf{A}$  is a probability measure on  $Y$ . Given a random variable (or equivalently, measurable function)  $T : Y \rightarrow Z$ ,  $\mathbf{A}T$  is a measurable function  $X \rightarrow Z$ . Given  $\mathbf{A} : X \rightarrow \Delta(Y)$  and  $\mathbf{B} : Y \rightarrow \Delta(Z)$ ,  $\mathbf{A}\mathbf{B}$  is a stochastic map  $X \rightarrow \Delta(Z)$ . We can use this same notation for continuous sets  $X$  and  $Y$ , see Çinlar [2011].

Write  $\mathbf{A}_x$  for the probability measure given by  $\delta_x\mathbf{A}$ , for  $E \subset X$  write  $\mathbf{A}_E$  for  $\mathbf{A}\mathbb{1}_E$  where  $\mathbb{1}_E$  is the indicator function on  $E$ , and write  $\mathbf{A}(x; E) := A_x(E) := \delta_x A_E$ . The tensor product  $\mathbf{A} \otimes \mathbf{B}$  is the stochastic map  $X \times Y \rightarrow \Delta(Y \times Z)$  given by  $(x, y) \mapsto \mathbf{A}_x\mathbf{B}_y$ .

Product notation is useful for defining composite kernels and probability measures, but sometimes more elaborate constructions are called for. Here we use string diagrams. String diagrams can always be interpreted as a mixture of matrix products and tensor products of Markov kernels, but we introduce kernels with special notation that helps with interpreting the resulting objects. A kernel  $\mathbf{A} : X \rightarrow \Delta(\mathcal{Y})$  is written  $\text{---}\mathbf{A}\text{---}$ , where the input and output wires are associated with the measurable spaces  $(X, \mathcal{X})$  and  $(Y, \mathcal{Y})$ . Probability measures  $\mu \in \Delta(\mathcal{X})$  are written  $\text{---}\mu\text{---}$  and measurable functions  $X \rightarrow Y$  are written  $\text{---}Y\text{---}$ . For a thorough definition of string diagrams, see Cho and Jacobs [2019].

The identity  $\text{Id}_X : X \rightarrow \Delta(X)$  is the Markov kernel  $x \mapsto \delta_x$ , which we represent with a bare wire, leaving the space implicit. The copy map  $\gamma : X \rightarrow \Delta(X \times X)$  is the Markov kernel  $x \mapsto \delta_{(x,x)}$ . For  $\mathbf{A} : X \rightarrow \Delta(Y)$  and  $\mathbf{B} : X \rightarrow \Delta(Z)$ ,  $\gamma(\mathbf{A} \otimes \mathbf{B}) = \sum_{x \in X} A_x \otimes B_x$ . The discard map  $\text{!} : X \rightarrow \Delta(\{*\})$  is the Markov kernel  $X \rightarrow \{*\}$  given by  $x \mapsto \delta_*$ , where  $*$  is a one element set.

Given  $\mu \in \Delta(X)$ ,  $\mathbf{A} : X \rightarrow \Delta(Y)$  as before, the joint distribution on  $X \times Y$  that might be informally written  $P(X)P(Y|X)$  is given in string diagram notation as

$$\triangleleft \mu \text{---} X \text{---} Y := \triangleleft \mu \text{---} X \text{---} \boxed{\mathbf{A}} \text{---} Y \quad (1)$$

A string diagram such as 1 that is “capped” on the left by a probability measure defines a probability space where the sample space is the Cartesian product of the output wires, the measurable sets are the tensor product of the output  $\sigma$ -algebras and the probability measure is given by the composition of measures in the diagram. The projection map  $\pi_X : X \times Y \rightarrow X$  is thus a measurable function; following this observation, we overload the notation for the random variable  $X$  to label wires on the diagram; when used as such, it always refers to the projection map  $\pi_X$ . While a random variable technically requires a probability space, we also use this convention for string diagrams representing kernels  $X \rightarrow \Delta(\mathcal{Y})$  for arbitrary  $X, Y$  (such diagrams feature “free” wires on the left and right). Using this convention, the measurable function referred to by the wire label  $X$  is always unambiguous, but we need to define a prior  $\xi \in \Delta(\mathcal{X})$  in order for it to have a distribution.

Finally, if we are given a set of kernels  $\{\mathbf{A}, \mathbf{B}\}$  where  $\mathbf{A} : X \rightarrow \Delta(\mathcal{Y})$ ,  $\mathbf{B} : W \rightarrow \Delta(\mathcal{Z})$  and a composition defining some kernel  $\mathbf{K} : T \rightarrow \Delta(\mathcal{U})$  where  $T, U$  are each Cartesian products of some subset of  $\{W, X, Y, Z\}$ , we can always construct  $\mathbf{K}^* : T \rightarrow \Delta(\mathcal{W} \otimes \mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{Z})$  by inserting copy maps in appropriate places. Thus if we have *both* a set of kernels and a diagram defining their composition, we can cautiously regard the input and output wires of *each* kernel as a random variable under the same interpretation as given above.

If we regard  $\nu$  as a joint distribution of  $X$  and  $Y$ , the marginal  $\nu_X$ , which by definition of  $\nu$  is equal to  $\mu$ , is given by

$$\triangleleft \nu \text{---} X \text{---} Y = \triangleleft \mu \text{---} X \quad (2)$$

A disintegration  $\nu_{Y|X}$  is any kernel  $X \rightarrow \Delta(\mathcal{Y})$  such that

$$\triangleleft \nu \text{---} X \text{---} Y = \triangleleft \nu_{Y|X} \text{---} Y := \triangleleft \nu \text{---} X \quad (3)$$

Disintegrations are known to exist wherever  $X$  and  $Y$  are standard measurable spaces (isomorphic to a discrete space or the reals with the Borel  $\sigma$ -algebra), though in general they are not unique. We will use the notation  $\nu_{Y|X}$  to refer to an arbitrary representative of the full set of disintegrations. Note that from Equations 1 and 2 it is clear that  $\mathbf{A}$  is a disintegration  $\nu_{Y|X}$ . We use disintegrations to represent conditional probability.

The copy map and erase maps have the following properties:

$$(4)$$

$$(5)$$

And for all  $\mathbf{A}$  with appropriate signature

$$(6)$$

A subset of the notation we use for string diagrams is the subject of a coherence theorem: taking a string diagram and applying a planar deformation or any of the rules 4, 5 and 6 yields a string diagram that represents the same kernel [Selinger, 2010]. This is a key reason for the power of string diagrams.

### 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”  $\Theta$ , a set of decisions  $D$  and a loss function  $l : \Theta \times D \rightarrow \mathbb{R}$ . 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  $H : \theta \mapsto \mu_\theta$  a *statistical experiment*. Given a *decision strategy*  $\pi : E \rightarrow \Delta(\mathcal{D})$ , define the *risk* of  $\pi$  given state  $\theta$  to be the expected loss of  $\pi$  in state  $\theta$ . Specifically,  $R : \Pi \times \Theta \rightarrow \mathbb{R}$  given by  $R : (\pi, \theta) \mapsto \delta_\theta \vee (H\pi \otimes \text{Id}_\Theta)l$ , where we make use of the product notation and copy map for brevity.

Supposing some unknown true state  $\theta^*$ , we would ideally find a strategy  $\pi$  that minimises the risk in  $\theta^*$ . Unfortunately, most statistical decision problems do not admit such strategies. Two alternative decision rules are available:

Given a measure  $\xi \in \Delta(\Theta)$  called a prior,  $\xi$ -*Bayes decision rule* is a decision rule  $\pi_{\text{Ba}}^*$  such that the *Bayes risk*  $R_\xi : \pi \mapsto \xi \vee (H\pi \otimes \text{Id}_\Theta)l$  is minimised:

$$\pi_{\text{Ba}}^* \in \arg \min_{\pi \in \Pi} R_\xi(\pi) \quad (7)$$

A *minimax* decision rule  $\pi_{\text{MM}}^*$  minimises the worst-case risk. Unlike a Bayes rule, it does not invoke a prior:

$$\pi_{\text{MM}}^* \in \arg \min_{\pi \in \Pi} \max_{\theta \in \Theta} R(\theta, \pi) \quad (8)$$

We emphasise here that we regard the set  $\Theta$  as a “state of nature” or a “theory of nature” and not a “parameter set” - it is possible that for some  $\theta \neq \theta'$  we have  $\mu_\theta = \mu_{\theta'}$ , a possibility not supported by the interpretation of  $\Theta$  as a set of distribution parameters. 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.

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  $\langle \Theta, E, D, H, l \rangle$  where  $\Theta, E$  and  $D$  are measurable sets,  $H$  is a stochastic map  $\Theta \rightarrow \Delta(\mathcal{E})$  and  $l$  a measurable function  $E \rightarrow \mathbb{R}$ . We leave implicit the set  $\Pi$  of decision strategies  $E \rightarrow \Delta(\mathcal{D})$  and  $\mathbb{R}$ , the codomain of  $l$ .

This is a very bare bones exposition of the theory of SDPs, and for more details we refer readers to Ferguson [1967].

Need a canonical measure on  $\Theta$ ; the coarsest measure rendering the evaluation maps measurable?

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

- Arbitrary sets of CBNs (e.g. ADMGs)

this might be equivalent to the set of 1-combs, see Jacobs

- Marginal models (e.g. mDAGs)
- CBNs with different intervention rules
- Cyclic SEMs (???)

don't know about this one

Show this for all the examples claimed

It is not the case, however, that *every* type of “causal graphical model” is a representation of a causal theory - SWIGs are an important exception. This is because SWIGs, like all counterfactual models, represent hypothesis classes (a hypothesis class is the range of a statistical experiment).

## 4 Causal Bayesian Networks

Suppose we have a set of “interventions”  $R$  which factorises as  $R = \otimes_{i \in [n]} \{\#\} \cup X^i$  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  $\{X^i | i \in \mathbb{N}\}$  such that  $X^i : E \rightarrow X^i$ . We denote an element  $(x^0, \#, \dots, \#, x^n) \in R$ ,  $x^0, x^n \neq \#$  by the notation  $do(X^0 = x^0, X^n = x^n)$  where occurrences of the distinguished element  $*$  are omitted. Denote by  $\underline{\#}$  the element of  $R$  consisting entirely of  $\#$  (equivalently,  $do()$ ).

For  $n \in \mathbb{N}$ , 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  $\{X^i\}_{i \in [n]}$  of random variables. That is, we have bijective functions  $f : V \rightarrow [n]$  and  $g : \{X^i\}_{i \in [n]} \rightarrow [n]$  and we adopt the convention that  $f(i) := V^i$  and  $g(i) := X^i$ . In addition, we will sometimes let a set  $U \subset V$  or  $a \subset [n]$  to denote a set of random variables rather than vertices or natural numbers; this is licenced by the bijections  $f$  and  $g$ . We will therefore overload notation and simply refer the nodes of  $\mathcal{G}$  as the random variables  $X^i$ .

We also suppose we have surjective  $h : R \rightarrow \mathcal{P}([n])$  such that  $h : (x^0, \dots, x^n) \mapsto \{i | x^i \neq *\}$ . That is,  $h$  picks out the indices that aren't suppressed in the  $do(\dots)$  notation for elements of  $V$ . Define  $X^{i'} : R \rightarrow \{\#\} \cup X^i$  by the function returning the  $i$ -th element of  $r$  for  $r \in R$ . Again, we suppose we have a bijection between primed random variables and natural numbers and can therefore pick out corresponding sets of primed RVs and unprimed RVs or natural numbers.

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

1.  $P_r$  is compatible with  $\mathcal{G}$  (see Pearl [2009])
2. For all  $i \in h(r)$ ,  $P_r F_{X^i} = \delta_{X^{i'}(r)}$
3. For all  $i \notin h(r)$ ,  $P_r |_{\text{Pa}_{\mathcal{G}}(X^i)} F_{X^i} = P_{\underline{\#}} |_{\text{Pa}_{\mathcal{G}}(X^i)} F_{X^i}$ ,  $P_{\underline{\#}}$ -almost surely

This definition differs slightly from that given in Pearl [2009]; for example  $P_*$  is a map to  $\Delta(\mathcal{E})$  rather than a set of labeled members of  $\Delta(\mathcal{E})$ , and we formulate it in directly in terms of measure theoretic probability rather than elementary probability. Nonetheless, we claim these choices don't meaningfully alter the standard definition, at least if we restrict  $E$  to be finite, and they make for a more convenient connection with CSDPs.

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

At least in the case of discrete  $E$  and  $P_*(\underline{\#})$  positive definite, we have from this definition for any  $r \in V$  the *truncated factorisation* property:

$$P_r F_{\mathbf{X}}(A) = \prod_{i \in h(r)} \delta_{X^{i'}(r)}(X^i(A)) \sum_{a \in A} \prod_{i \notin h(r)} P_{\underline{\#}} |_{\text{Pa}_{\mathcal{G}}(X^i)} F_{X^i}(a; \{X^i(a)\}) \quad (10)$$

I haven't found a formulation of CBNs on infinite spaces, let alone continuous ones



As a consequence of the existence of conditional probability, given  $\mathcal{G}$  and  $\mu$  there exists a unique set of interventional maps  $P_*$  compatible with both  $\mathcal{G}$  and  $\mu$  as above. This property licenses a typical use case of CBNs:  $\mathcal{G}(\cdot)$  is treated as a *map* from the subset of  $\Delta(\mathcal{E})$  compatible with  $\mathcal{G}$  to interventional maps  $V \rightarrow \Delta(\mathcal{E})$ . More generally, provided  $\mu$  is compatible with  $\mathcal{G}$  we have that  $\text{Io}$  exists, and so  $\mathcal{G}(\mu)$  is non-empty.

Condition 3 presents some difficulties in the presence of measure 0 sets, as when a conditional probability such as  $P_{\#|\text{Pa}_{\mathcal{G}}(V^i)}$  may be variously intended to mean a particular element of the class of conditional probabilities, any element or every element in the class class [Çinlar, 2011]; condition 3 will have different implications for these various interpretations.

Letting  $\mathcal{H}^{\mathcal{G}} \subset \Delta(\mathcal{E})$  be some *hypothesis class* of probability measures compatible with a causal graph  $\mathcal{G}$ , define the set of pairs  $\mathcal{T}^{\mathcal{G}} := \{(\mu, \kappa) | \mu \in \mathcal{H}^{\mathcal{G}}, \kappa \in \mathcal{G}(\mu)\}$ . Recall that a causal theory can be represented as a set of (observation, consequence map) pairs; i.e.  $\mathcal{T}^{\mathcal{G}}$  is a causal theory. The map  $\text{Th} : \mathcal{G} \mapsto \mathcal{T}^{\mathcal{G}}$  is therefore a map from directed acyclic graphs to causal theories. Unlike the map from DAGs to sets of probability measures, the map from DAGs to causal theories is injective.

**Theorem 4.2** (The map  $\text{Th}$  is injective). *For DAGs  $\mathcal{G}, \mathcal{G}'$  on the same set of RV's  $\{X^i\}_{i \in [n]}$ ,  $\mathcal{G} \neq \mathcal{G}' \implies \mathcal{T}^{\mathcal{G}} \neq \mathcal{T}^{\mathcal{G}'}$ .*

*Proof.* Sketch:  $\mathcal{G}$  and  $\mathcal{G}'$  must disagree on at least one parental set. Choose some  $\mu$  such that  $P_{\#|\text{Pa}_{\mathcal{G}}(X^i)} F_{X^i} \neq P_{\#|\text{Pa}_{\mathcal{G}'}(X^i)} F_{X^i}$ . Then take  $r, r'$  such that  $h(r) = h(r') = \text{Pa}_{\mathcal{G}}(X^i) \cup \text{Pa}_{\mathcal{G}'}(X^i)$ ,  $\text{Pa}_{\mathcal{G}}(X^i)(r) = \text{Pa}_{\mathcal{G}}(X^i)(r')$  but  $r \neq r'$ . Then  $P_r F_{X^i} \neq P_{r'} F_{X^i}$  so  $P_r \neq P_{r'}$ .  $\square$

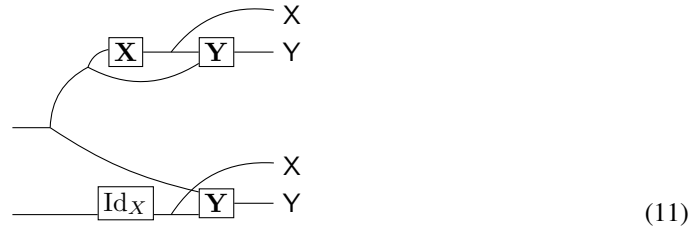
Each DAG  $\mathcal{G}$  represents a causal theory  $\mathcal{T}^{\mathcal{G}}$ . For every causal theory  $\mathcal{T}$ , either it is not represented by any graph or there is a unique graph  $\mathcal{G}$  such that  $\mathcal{T} = \mathcal{T}^{\mathcal{G}}$ . It is in this sense that we claim Causal Bayesian Networks are a subset of causal theories.

The string diagram notation we use to represent Markov kernels and the DAGs used to represent CBNs have clear similarities. Fong [2013] discusses how a DAG can be translated to a string diagram to yield a different type of “causal theory”. It is in fact possible to represent the causal theory associated with what we call an elementary CBN compactly in string diagram notation. An elementary CBN is a CBN where only one node accepts intervention and the “do-nothing” action is not available. We define it directly as a causal theory. Rather than formally define how this translation may proceed we will present an example demonstrating how this is possible.

**Definition 4.3** (Elementary Causal Bayesian Network). Given  $D, E, \Theta$ , random variables  $\{X^i\}_{i \in [n]}$  on  $E$ , a distinguished variable  $X^0$  taking values in  $D$  and a causal theory  $T : \Theta \times D \rightarrow \Delta(\mathcal{E} \otimes \mathcal{E})$  with  $H := T(\text{Id}_E \otimes *_E)$  and  $C := T(*_E \otimes \text{Id}_E)$ , an *elementary Causal Bayesian Network* (eCBN) compatible with  $T$  is a directed acyclic graph (DAG)  $\mathcal{G}$  with nodes  $\{X^i\}_{i \in [n]}$  such that

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

Suppose we have the EDAG  $\mathcal{G} := X \rightarrow Y$ , 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  $T : \Theta \times X \rightarrow \Delta([\mathcal{X} \otimes \mathcal{Y}]^2)$  if and only if there exist Markov kernels  $\mathbf{X} : \Theta \rightarrow \Delta(\mathcal{X})$ ,  $\mathbf{Y} : \Theta \times X \rightarrow \Delta(\mathcal{Y})$  such that



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.

General definition:  $f(A)$  is the image of  $A$  under  $f$  and  $X$  as the copy-mapped tensor product of vari-

Surely Pearl or a student has dealt with this somewhere? Any element seems to be the most appropriate choice, but this renders CBNs useless for continuous spaces unless we place extra restrictions on  $P_*$

*Proof.* (Sketch): The topology of the top and bottom sub-structures guarantees 1 (compatibility) and 1 guarantees that some kernels exist exhibiting this topology (this condition is actually trivial in this case; it is nontrivial where the graph is not fully connected). 2 is equivalent to asserting that  $C_{\theta,d}F_{X^i}$  is the identity map. The shared kernel  $\mathbf{Y}$  guarantees 3 and if  $\mathbf{Y}$  cannot be shared then 3 does not hold.  $\square$

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.

We can't avoid one condition being trivial with two nodes, but three nodes starts looking very complex!

## 5 Potential outcomes models

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

Notationally, we will refer to the symbol  $W_i$  as the  $i$ -th treatment assignment ( $i \in \{0, \dots, n\}$ ),  $Y_i(0)$   $Y_i(1)$  as the  $i$ -th potential outcomes,  $Y_i$  as the  $i$ -th observed outcome and  $X_i$  as the  $i$ -th “vector of background facts”.  $W$  refers to the bundle of all  $W_i$ s and similarly for other symbols. Following the convention set out in our introduction, we consider these symbols to represent random variables and to label strings in our string diagram. Suppose the vector  $[Y_0, \dots, Y_n]$  takes values in  $Y$  and similarly for other symbols.

Given an underlying state space  $\Theta$ , 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  $\mathbf{H} : \Theta \rightarrow \Delta(\mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{W})$ .

The kernels are

- A “model on the science”,  $\mathbf{P} : \Theta \rightarrow \Delta(\mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{Y})$  (In Rubin’s notation,  $\mathbf{P}$  is  $\prod_i f(X_i, Y_i(0), Y_i(1)|\theta)$ )
- An “assignment mechanism”,  $\mathbf{W} : \Theta \times X \times Y^2 \rightarrow \Delta(\{0, 1\}^n)$  (in Rubin’s notation,  $\mathbf{W}$  is  $Pr(W|X, Y(1), Y(0))$ )
- An “observation model”,  $\mathbf{Y} : \{0, 1\}^n \times Y^2 \rightarrow \Delta(\mathcal{Y})$ , defined explicitly as  $\mathbf{Y} : (y^0, y^1, \mathbf{w}) \mapsto (1 - \mathbf{w}) \odot \delta_{y^0} + \mathbf{w} \odot \delta_{y^1}$  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  $W_i = w \implies Y_i = Y_i(w)$  *almost surely* in some probability space, as a Markov kernel cannot guarantee exact equality. We also differ from Rubin by including  $\Theta$  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  $\Theta$  is larger than Rubin’s).

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

$$\mathbf{H}^\mathcal{O} := \text{Diagram} \quad (12)$$

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

### 5.1 Are potential outcomes models causal theories?

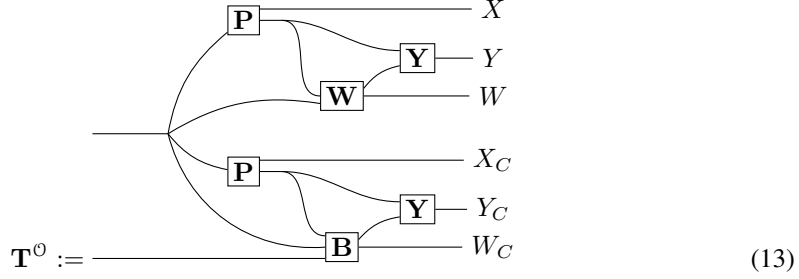
By assumption, a potential outcomes model induces a canonical statistical experiment. Given that potential outcomes models are a type of causal model, we can ask whether they induce a canonical



*causal theory*. We propose one strategy for constructing such a theory below, though it is known to work only where all associated spaces are discrete.

Given a potential outcomes model  $\mathcal{O} := \langle \mathbf{P}, \mathbf{W}, \mathbf{Y} \rangle$  with associated discrete spaces  $\Theta, X, Y, [m]$ , define  $D := [0, 1]^{|\Theta|+2|Y|+|X|}$ . Then there is a kernel  $\mathbf{B} : D \times \Theta \times X \times Y^2 \rightarrow \Delta(\mathcal{W})$  such that for every Markov kernel  $\mathbf{W}' : \Theta \times X \times Y^m \rightarrow \Delta([m]^n)$  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*  $\mathbf{T}^\mathcal{O}$ :



$\mathbf{T}^\mathcal{O}$  is two parallel copies of  $\mathcal{H}^\mathcal{O}$  where  $\mathbf{W}$  is replaced by  $\mathbf{B}$  in the lower version. We justify the claim that is is appropriate to consider  $\mathbf{T}^\mathcal{O}$  the causal theory associated with  $\mathcal{O}$  on the basis of two considerations: firstly, in practice decisions are often considered to cause modifications of the treatment assignment function  $\mathbf{W}$ . Secondly, the map  $\mathcal{O} \mapsto \mathbf{T}^\mathcal{O}$  identifies two potential outcomes models  $\mathcal{O}$  and  $\mathcal{O}'$  if and only if they feature the same “science”  $\mathbf{P}$  and induce the same statistical experiment  $\mathbf{H}$ .

**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  $\mathbf{C} := \mathbf{T}(* \otimes \text{Id}) = \mathbf{T}'(* \otimes \text{Id})$  as all kernels in the bottom half of 13 ( $\mathbf{P}, \mathbf{B}$  and  $\mathbf{Y}$ ) are the same by definition. But then  $\mathbf{T} = \vee(\mathbf{H} \otimes \mathbf{C}) = \mathbf{T}'$ .

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  $\mathbf{T}_{\theta,d}^\mathcal{O} \pi_{\mathcal{W}}(\{0\}) = \mathbf{P}_\theta(A) \neq \mathbf{P}'_\theta(A) = \mathbf{T}_{\theta,d}^{\mathcal{O}'} \pi_{\mathcal{W}}(\{0\})$ , a contradiction. Thus  $\mathbf{P} = \mathbf{P}'$ . In addition,  $\mathbf{H} = \mathbf{T}(\text{Id} \otimes *) = \mathbf{H}'$ .  $\square$

Note that the assignment  $\mathbf{W}$  may differ between  $\mathcal{O}$  and  $\mathcal{O}'$ . Suppose  $X = \emptyset$ ,  $Y = \{0, 1\}$  and for some  $\theta$ ,  $\mathbf{P}_\theta = \frac{1}{4}(\delta_{0,0} + \delta_{0,1} + \delta_{1,0} + \delta_{1,1})$ . Then  $\mathbf{W}_\theta : (y_0, y_1) \mapsto \llbracket y_0 = y_1 \rrbracket \delta_0 + \llbracket y_0 \neq y_1 \rrbracket \delta_1$  and  $\mathbf{W}'_\theta := 1 - \mathbf{W}_\theta$  both yield the same observations  $\mathbf{H}_\theta$ .

The causal theory  $\mathbf{T}^\mathcal{O}$  is an unrealistic theory. We usually do not have at our disposal a set of decisions that can induce any possible treatment assignment function. Under  $\mathbf{T}^\mathcal{O}$  we have the possibility (among others) of deciding to assign treatment if and only if  $y_1 > y_0$ . Note that this appears to have something in common with CBNs: both feature a causal theory with unreasonably many actions, and in fact both include the possibility of decisions that render the problem trivial. We will call such theories that posit unrealistic amounts of control *dextrous theories*.

In order to choose a decision, we want to work with a *pragmatic theory*  $\mathbf{T}^p$  that describes a more restricted set of decisions  $D^p$  that correspond to those actions we believe we can actually take. However, it might be the case that we are willing to believe that all pragmatic decisions  $d_p \in D^p$  have the same consequences as corresponding decisions  $f(d_p) \in D$  in the dextrous theory.

To illustrate this, consider the problem of evaluating the “effect of assigning treatment” vs “the effect of receiving treatment” (the former being known as *intention to treat analysis*). 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 of receiving 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 deterministically setting  $W = 1$  and  $W = 0$  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  $e \in D^{\text{ITT}}$  such that for all  $\theta, x, y_0, y_1$ ,  $\mathbf{B}_{\theta, d, x, y_0, y_1}^{\text{ITT}} = \delta_1$ .

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  $D^{\text{RT}}$  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. That is, we have a vague idea that there might be a correspondence between  $D^{\text{RT}}$  and  $D^p$ , and we could express our uncertainty with a set of possible correspondences or a probability measure over correspondences, but we don’t clearly have a single correspondence to work with.

Finally, 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  $W^{\text{ITT}}$  is consistent with the test subjects always taking the treatment. That is, we’re not just uncertain about the correspondence between dextrous decisions  $D^{\text{ITT}}$  and pragmatic decisions  $D^t$  – we are in fact confident that there is no such correspondence.

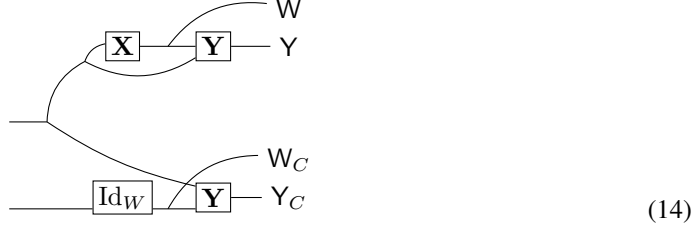
## 6 Comparing Causal Bayesian Networks and Potential Outcomes theories

We can easily see that the causal theories associated with a CBN and a potential outcomes model are typically 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.

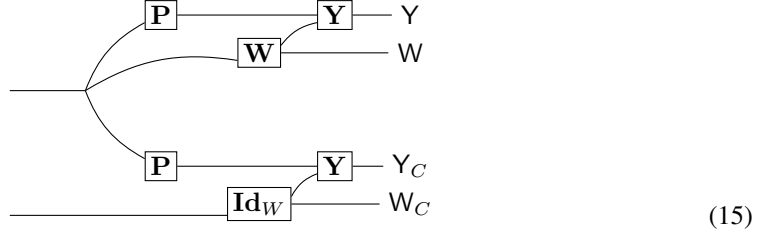
However, it may be the case that the pragmatic theory we derive from a dextrous CBN or PO theory is common to both.

Suppose we have a CBN  $\mathcal{G} := W \rightarrow Y$ , where  $W$  and  $Y$  are random variables taking values in some arbitrary spaces  $W$  and  $Y$ . Suppose also that we require a pragmatic theory  $\mathbf{T}^{\mathcal{G}} : \Theta \times W \rightarrow \Delta([W \otimes \mathcal{Y}]^2)$  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  $\mathbf{W} : \Theta \rightarrow \Delta(W)$ ,  $\mathbf{Y} : \Theta \times W \rightarrow \Delta(\mathcal{Y})$  such that  $\mathbf{T}^{\mathcal{G}}$  can be represented as in the diagram 14. Conversely, any causal theory that can be represented in this

manner is a candidate for  $\mathbf{T}^G$  (see A)



Suppose we have a PO model  $\mathcal{O} = \langle \mathbf{P}, \mathbf{W}, \mathbf{Y} \rangle$  on  $\Theta$ ,  $W$  and  $Y$  such that  $\mathbf{W}$  depends only on  $\Theta$ . 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 15. Conversely, there is a PO model for every causal theory with this representation.



## 7 No causes in no causes out

A key result in statistical learning theory is the requirement that, in order for a hypothesis class to be learnable, it must have finite VC-dimension. The concept of controlling the size of the hypothesis class plays a fundamental role across the field of machine learning, from formal proofs of learnability to techniques based less formally on the notion of the bias-variance tradeoff. CSDPs are closely related to statistical learning problems, and it is highly likely that results of this type can be developed for causal problems.

Apart from any inductive biases necessary for learnability, causal theories also require a *decision bias* - a causal theory that does not distinguish decisions yields only trivial results. This is distinct from a restriction on the flexibility or capacity of a causal theory. Given a prior, the requirement is that, conditional on some set of observations, a causal theory yields different consequences for different decisions.

Define the pairwise swap  $U_{dd'} : D \rightarrow \Delta(\mathcal{D})$  to be the kernel that sends  $d \mapsto \delta_{d'}$ ,  $d' \mapsto \delta_d$  and all other  $d'' \mapsto \delta_{d''}$ .

**Theorem 7.1** (No causes in, no causes out (Bayes)). *If a causal theory  $T : \Theta \times D \rightarrow \Delta(\mathcal{E} \otimes \mathcal{F})$  and a prior  $\xi \in \Delta(\Theta)$  are such that for all pairwise swaps  $U_{dd'} : D \rightarrow \Delta(\mathcal{D})$ ,  $(\xi \otimes U_{dd'})T = (\xi \otimes I)T$  and  $D$  is discrete then all decision strategies are Bayes.*

*Proof.* Defining  $F_{-d_0} : d \mapsto \delta_{d_0}$  for all  $d \in D$ , we will show that for all  $J$ ,  $S_\xi(J) = S_\xi(JF_{-d_0}) := S_0$ .

By assumption, for all  $d \in D$ , utility functions  $u$ :

$$\int_{\Theta} H_{\theta} J(\{d\}) C_{\theta} u(d) d\xi = \int_{\Theta} H_{\theta} J(\{d\}) U_{dd_0} C_{\theta} u(d) d\xi \quad (16)$$

$$= \int_{\Theta} H_{\theta} J(\{d\}) F_{-d_0} C_{\theta} u(d) d\xi \quad (17)$$

$$\therefore \sum_{d \in D} \int_{\Theta} H_{\theta} J(\{d\}) F_{-d_0} C_{\theta}(d; A) d\xi = \sum_{d \in D} \int_{\Theta} H_{\theta} J(\{d\}) C_{\theta} u(d) d\xi \quad (18)$$

$$= \int_{\Theta} \sum_{d \in D} H_{\theta} J(\{d\}) C_{\theta} u(d) d\xi \quad (19)$$

$$= \int_{\Theta} H_{\theta} J C_{\theta} u d\xi \quad (20)$$

$$= S_{\xi}(J) \quad (21)$$

$$= S_{\xi}(J F_{-d_0}) \quad (22)$$

Where 17 follows from the fact that evaluation at  $d$  guarantees  $U_{dd_0} C_{\theta} u(d) = F_{-d_0} C_{\theta} u(d)$ .  $\square$

**Corollary 7.2.** *If a causal theory  $T$  with a prior  $\xi$  and discrete decision set  $D$  yields a nontrivial ordering of decision strategies, then there exists  $d, d' \in D$  such that  $(\xi \otimes \delta_d)T \neq (\xi \otimes \delta_{d'})T$ .*

Somewhat surprisingly, the minimax rule may yield preferences over decisions under such circumstances; in particular, a uniform strategy is always minimax, though other strategies may not be. This is because the consequences of a uniform strategy may be less extreme than the consequences of any other strategy.

**Theorem 7.3** (No causes in, uniform strategy out (minimax)). *If a causal theory  $T : \Theta \times D \rightarrow \Delta(\mathcal{E} \otimes \mathcal{F})$  with finite  $D$  is such that for all pairwise swaps  $U_{dd'} : D \rightarrow \Delta(\mathcal{D})$ ,  $\theta \in \Theta$  there is some  $\theta'$  such that  $T_{\theta, \cdot} = (I \otimes U)T_{\theta', \cdot}$ , then the uniform decision strategy is minimax.*

*Proof.* Note that for finite  $D$ , the invertible maps  $D \rightarrow \Delta(\mathcal{D})$  are permutation maps which can be factorised as a sequence of pairwise swaps.

Call  $J_U$  the stubborn uniform strategy  $J_U : x \mapsto U(\mathcal{D})$  for all  $x \in E$ . Suppose there is some nonuniform  $J$  such that  $\max_{\theta} S(J, \theta) < \max_{\theta} S(J_U, \theta)$ . Suppose  $S(J_U, \theta)$  is maximised in some state  $\theta^0$  where  $S(J_d, \theta^0) = S(J_{d'}, \theta^0)$  for all  $d, d' \in D$ . Then  $S(J, \theta^0) = S(J_U, \theta^0)$ , contradicting our assumption that  $J$  achieved lower risk in the worst case. Suppose  $S(J_U, \theta)$  is maximised in some state  $\theta^1$  where there are some  $d, d' \in D$  such that  $S(J_d, \theta^1) > S(J_{d'}, \theta^1)$ . Then there are most  $|D|/2$  decisions where  $S(J_d, \theta^1)$  is greater than the median of  $A = \{S(J_d, \theta^1) | d \in D\}$  and at least one such decision, and at least  $|D|/2$  decisions such that  $\mu_{\theta^1} J(d)$  is greater than or equal to the median of  $B = \{\mu_{\theta^1} J(d) | d \in D\}$ , with at least one strictly greater. Thus there is an invertible map  $f : D \rightarrow D$  such that  $f(A) \subset B$ . But then there is some  $\theta^2$  such that  $S(J_d, \theta^1) = S(J_{f(d)}, \theta^2)$  for all  $d \in D$  and thus  $S(J, \theta^2) > S(J_U, \theta^2) = S(J_U, \theta^1)$  contradicting our assumption that  $J$  was better by the minimax rule than  $J_U$ .  $\square$

**Corollary 7.4.** *If the risk of the uniform strategy is maximised in some state  $\theta^*$  such that  $S(J_d, \theta^*) > S(J_{d'}, \theta^*)$  for some  $d, d'$ , then the uniform strategy is strictly better than any nonuniform strategy.*

Thus for a causal theory to support nontrivial results, we require for Bayes rules a prior  $\xi$  such that  $(\xi \otimes \delta_d)T$  depends on  $d$ , or for the minimax rule that the set of distributions mapped by the theory  $\mathcal{T}_d := \{T_{\theta, d} | \theta \in \Theta\}$  depends on the decision  $d$ . We will say that such theories/priors exhibit a *decision bias*.

From one point of view, this result might be expected: if we believe

- Any possible consequence of  $d_1$  might equally be a consequence of  $d_2$  and vice versa
- Any data we encounter is equally consistent with  $d_1$  having some set of consequences or with  $d_2$  having that same set of consequences



In other words, if  $T$  can be extended to  $T'$  via  $M$ , then we can “save” the results of conditioning  $T_\xi$  on  $A$  via  $T_\xi|A$  and later on we can determine the effects of some strategy  $\gamma$  with respect to  $T'$  via  $\gamma MT_\xi|A$ .

We will return to our discussion of the the “effect of taking the treatment” for an example. Suppose we have  $\Theta = [0, 1]^2 := \Theta_1 \otimes \Theta_2$  where given  $(\theta_1, \theta_2) \in \Theta$  we identify  $\theta_1$  with “treatment efficacy” and  $\theta_2$  with “treatment susceptibility”. Let  $Y, W = \{0, 1\}$  where  $Y$  is the set of outcomes and  $W$  indicates whether or not a patient took the treatment. Define a potential outcomes model  $H_{PO} : \Theta \rightarrow \Delta(\mathcal{Y}^2)$ ,  $H_W : \Theta \rightarrow \Delta(\mathcal{W})$  and  $H_Y : W \times Y^2 \rightarrow \Delta(\mathcal{Y})$ . Furthermore, suppose we have  $D = [0, 1]^2$  and  $C_W : \Theta_2 \times D \rightarrow \Delta(\mathcal{W})$  defined by  $C_W(\theta_2, d_1, d_2; A) := \theta_2(d_1\delta_1(A)(1 - d_1)\delta_0(A)) + (1 - \theta_2)(d_2\delta_1(A) + (1 - d_2)\delta_0(A))$ ; that is,  $d_1$  and  $d_2$  parametrise the set of Markov kernels  $\Theta_2 \rightarrow \Delta(\mathcal{W})$ . Then  $\langle H_{PO}, H_W, H_Y, C_W \rangle$  defines a causal theory  $T$ . Suppose we observe  $A \in \mathcal{E}$ ; then for  $(d_1, d_2) \in D$ ,  $B \in \mathcal{E}$ :

$$T_\xi|A(d_1, d_2; B) = \frac{1}{\int_{\Theta} (H_{PO,\theta} \otimes H_{W,\theta}) H_Y(A) C_{W,\theta}(d_1, d_2; B) d\xi} \int_{\Theta} (H_{PO,\theta} \otimes H_{W,\theta}) H_Y(A) C_{W,\theta}(d_1, d_2; B) d\xi \quad (28)$$

$T_\xi|A$  describes a Markov kernel from  $D \rightarrow \Delta(\mathcal{E})$ . However, it is unlikely to be the case that  $D$  describes the actual decisions we have available - we probably don’t have the ability to choose the exact relationship between treatment susceptibility and treatment taking. In fact, all that may be actually available are some decisions  $D' = \{0, 1\}$  where 0 represents no prescription and 1 represents prescription; these may yield uncertain relationships between  $\theta_2$  and  $W$ . While we might not know which relationship a decision to prescribe or not induces, we might accept that influencing this relationship is the only important consequence of this decision. That is, we might accept that for *some*  $M : D' \rightarrow \Delta(\mathcal{D})$  a causal theory  $T' : \Theta \times D' \rightarrow \Delta(\mathcal{E}^2)$  of the form  $T' = (\text{Id}_\Theta \otimes M)T$  is appropriate. Then, because  $T'$  is an extension of  $T$ , rather than having to rerun our inference we can simply compute  $MT_\xi|A$ .

## References

- 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.
- 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-string-diagrams/0581C747DB5793756FE135C70B3B6D51>.
- 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.
- M. A. Hernán and S. L. Taubman. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity*, 32(S3):S8–S14, August 2008. ISSN 1476-5497. doi: 10.1038/ijo.2008.82. URL <https://www.nature.com/articles/ijo200882>.
- L. Le Cam. Comparison of Experiments - A Short Review.pdf. *IMS Lecture Notes - Monograph Series*, 30, 1996. URL <https://www.fastmailusercontent.com/mail-attachment/f2132276u33762/%3CB868E7FAD8F0C04FA0207113AD7C77A9%40ausprd01.prod.outlook.com%3E/Comparison%20of%20Experiments%20-%20A%20Short%20Review.pdf?u=9a2e41ab&a=195efd739077a3e70251567c111372cb67a6dcf0>.



- Jan Lemeire and Dominik Janzing. Replacing Causal Faithfulness with Algorithmic Independence of Conditionals. *Minds and Machines*, 23(2):227–249, May 2013. ISSN 0924-6495, 1572-8641. doi: 10.1007/s11023-012-9283-1. URL <https://link.springer.com/article/10.1007/s11023-012-9283-1>.
- David Lewis. Causal decision theory. *Australasian Journal of Philosophy*, 59(1):5–30, March 1981. ISSN 0004-8402. doi: 10.1080/00048408112340011. URL <https://doi.org/10.1080/00048408112340011>.
- 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.
- Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016. ISSN 1467-9868. doi: 10.1111/rssb.12167. URL <https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12167>.
- 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>.
- 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.
- 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.
- Peter Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.
- 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 Appendix: CBN representation as a causal theory

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

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