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 (**?**), who is also noted for his argument that knowledge of the latter cannot be reduced to the former. **?** 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 **?**). 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 **?**, underpins much of modern machine learning. It introduced the fundamental notions of *loss* and *risk* to statistics and proided 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 (**?**), SDT itself is regarded as a theory of statistical decision making and not of causality.

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

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/utililties of decisions, the latter "forgets about the consequences" and only works with reward/utiltiy.

Preprint. Under review.

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 $\{1, ..., n\}$. Sets are ordinary capital letters X, σ-algebras are script letters $\mathcal{X}$ while random variables are sans serif capitals $\mathsf{X} : \_ \to X$. 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 $\Delta(\mathcal{X})$ and the set of all stochastic maps $W \to \Delta(X)$ as $\Delta(\mathcal{X})^W$. $\delta_x : (X) \to [0, 1]$ is the probability measure such that $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise.

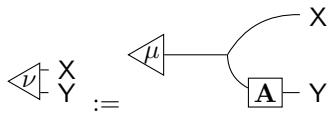
If X is a discrete space, probability measures on X are positive row vectors in $\mathbb{R}^{|X|}$ that sum to 1, and stochastic maps or Markov kernels $X \to \Delta(Y)$ are $|X| \times |Y|$ positive matrices with row sums of 1. Using the standard notion of associative matrix-vector products, given $\mu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

Write $\mathbf{A}_x$ for the probability measure given by $\delta_x\mathbf{A}$, for $E \subset X$ write $\mathbf{A}_E$ for $\mathbf{A}\mathbb{1}_C$ 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 $X \times Y \to \Delta(Y \times Z)$ 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 $\mathbf{A} : X \to \Delta(\mathcal{Y})$ is written ..., where the input and output wires are associated with the measureable spaces $(X, \mathcal{X})$ and $(Y, \mathcal{Y})$. Probability measures $\mu \in \Delta(\mathcal{X})$ are written ... and measurable functions $X \to Y$ are written .... For a thorough definition of string diagrams, see **?**.

The identity $\mathrm{Id}_X : X \to \Delta(X)$ is the Markov kernel $x \mapsto \delta_x$, which we represent with a bare wire, leaving the space implicit. 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 $\mathbf{B} : X \to \Delta(Z)$, $\curlyvee(A \otimes B) = \sum_{x \in X} A_x \otimes B_x$. The discard map $\top$ is the Markov kernel $X \to \{*\}$ given by $x \mapsto \delta_*$, where $*$ is a 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 as

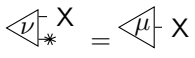


(1)

A string diagram such as **??** 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 σ-algebras and the probability measure is given by the composition of measures in the diagram. The projection map $\pi_X : X \times Y \to X$ 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 projetion map π_X. While a random variable technically requires a probability space, we also use this convention for string diagrams representing kernels $X \to \Delta(\mathcal{Y})$ 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.

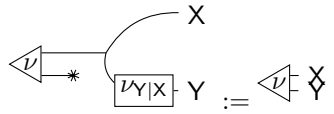
Finally, if we are given a set of kernels $\{\mathbf{A}, \mathbf{B}\}$ where $\mathbf{A} : X \to \Delta(\mathcal{Y})$, $\mathbf{B} : W \to \Delta(\mathcal{Z})$ and a composition defining some kernel $\mathbf{K} : T \to \Delta(\mathcal{U})$ where T, U are each Cartesian products of some subset of of $\{W, X, Y, Z\}$, we can always construct $\mathbf{K}^* : T \to \Delta(\mathcal{W} \otimes \mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{Z})$ 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 ν as a joint distribution of X and Y, the marginal ν_{X}, which by definition of ν is equal to μ, is given by



(2)

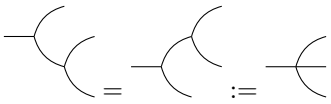
A disintegration $\nu_{\mathsf{Y}|\mathsf{X}}$ is any kernel $X \to \Delta(\mathcal{Y})$ such that



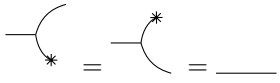
(3)

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

The copy map and erase maps have the following properties:

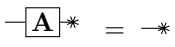


(4)



(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 **??**, **??** and **??** yields a string diagram that represents the same kernel (**?**). 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" 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 $H : \theta \mapsto \mu_\theta$ a *statistical experiment*. Given a *decision strategy* $\pi : E \to \Delta(\mathcal{D})$, define the *risk* of π given state θ to be the expected loss of π in state θ. Specifically, $R : \Pi \times \Theta \to \mathbb{R}$ given by $R : (\pi, \theta) \mapsto \delta_\theta \curlyvee (H\pi \otimes \mathrm{Id}_\Theta) l$, where we make use of the product notation and copy map for brevity.

Need a canonical measure on Θ; the coarsest measure rendering the evaluation maps measurable?

Supposing some unknown true state θ^*, we would ideally find a strategy π that minimises the risk in θ^*. 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, *ξ-Bayes decision rule* is a decision rule π^*_{Ba} such that the *Bayes risk* $R_\xi : \pi \mapsto \xi \curlyvee (H\pi \otimes \mathrm{Id}_\Theta) l$ is minimised:

$$\pi^*_{\mathrm{Ba}} \in \arg\min_{\pi \in \Pi} R_\xi(\pi) \tag{7}$$

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

$$\pi^*_{\mathrm{Mm}} \in \arg\min_{\pi \in \Pi} \max_{\theta \in \Theta} R(\theta, \pi) \tag{8}$$

We emphasise here that we regard the set Θ 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 Θ 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, **?**, 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.

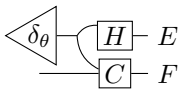
Formally, we define an SDP as the tuple $\langle \Theta, E, D, H, l \rangle$ where Θ, E and D are measurable sets, H is a stochastic map $\Theta \to \Delta(\mathcal{E})$ and l a measurable function $E \to \mathbb{R}$. We leave implicit the set Π of decision strategies $E \to \Delta(\mathcal{D})$ and $\mathbb{R}$, the codomain of l.

This is a very bare bones exposition of the theory of SDPs, and for more details we refer readers to **?**.

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. In many areas 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 $u : F \to \mathbb{R}$; "utility" being a very conventional term for such a desirability function. Given such a u, the tuple $\langle \Theta, E, D, H, u \rangle$ is an ill-posed problem; we want to evaluate the desirability of decisions D (or decision strategies π), but we have no means of connecting decisions with consequences F. An obvious move is to introduce for each state of nature θ a *consequence map* $\kappa_\theta : D \to \Delta(\mathcal{F})$; let C be the map $\theta \mapsto \kappa_\theta$. We can then define the *causal risk* $S : \Pi \times \Theta \to \mathbb{R}$ by $S : (\pi, \theta) \mapsto -\delta_\theta \curlyvee (H\pi \otimes ID\Theta)Cu$, and Bayes and minimax risks are defined by obvious analogy.

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

$T_\theta :=$ 

(9)

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

This is highly nonobvious, depends on the work of Jacobs, and is only known to be true for finite sets E and F. On the other hand, there's a very intuitive graphical proof.

Given a CSDP $\alpha = \langle \Theta, E, F, D, T, u \rangle$ where T is a theory arising from some H and C as in Equation **??**, we can recover $H = T(\mathrm{Id}_E \otimes *_F)$ and $C = T(*_E \otimes \mathrm{Id}_F)$. Given α and letting $l := Cu$ we induce the canonical SDP $\beta = \langle \Theta, E, D, H', l \rangle$ such that for any $\theta \in \Theta$, $\pi \in \Pi$, $R^{(\beta)}(\pi, \theta) = S^{(\alpha)}(\pi, \theta)$, and thus, if we accept that the risk functional is the only means of evaluating the desirability of a strategy (whether we choose Bayes, minimax or some other meta-rule to select a strategy), α and β will always produce identical recommendations.

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

Thus, in some sense every problem that can be represented as an SDP can be represented as a CSDP and vise-versa (this may not hold true if we invoke some decision rule that doesn't just depend on the risk functional).

We will finally note an alternative representation of a CSDP. Consider the set $\mathcal{T} = \{T_\theta | \theta \in \Theta\}$ and the evaulation map $\mathrm{Ev}_\mathcal{T} : \mathcal{T} \times D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ given by $(T_\theta, d) \mapsto T_\theta(d; \cdot)$. Then the problems $\alpha = \langle \Theta, E, F, D, T, u \rangle$ and $\alpha' = \langle \mathcal{T}, E, F, D, \mathrm{Ev}_\mathcal{T}, u \rangle$ are related in the sense that $S^{(\alpha)}(\pi, \theta) = S^{(\alpha')}(\pi, T_\theta)$. Thus, unlike with regular SDPs where Θ is an arbitrary set, for CSDPs we can regard Θ as a subset of kernels $D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ that can be written as in Eq. **??** with the canonical kernel $\mathrm{Ev}_\mathcal{T}$. We call $\mathcal{T}$ the *set causal theory* and, as there is a bijection between kernel theories T along with their domains Θ and set theories $\mathcal{T}$, we typically refer to either as simply a *causal theory*. We say that α is a CSDP in kernel form and α' is a CSDP in set form.

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

Finally, we note that causal theories can *also* be represented as a set of (distribution, consequence map) pairs.

Theorem 3.1 (Causal theories are sets of pairs)**.** *There is a bijection between the set of causal theories and the power set* $\mathcal{P}(\Delta(\mathcal{E}) \times \Delta(\mathcal{F})^D)$

Proof. Sketch: Theories to pairs goes $\{(T_\theta(\mathrm{Id}_E \otimes *_F), T_\theta(*_E \otimes \mathrm{Id}_F)) | T_\theta \in \mathcal{T}\}$ and pairs to theories goes $\{\mu \otimes \kappa | (\mu, \kappa) \in \mathrm{Pairs}\}$. □

A *causal theory* - that is a "consequence-aware" analogue of a statistical experiment. It can be represented as a Markov kernel or a set of (observation, consequence) pairs. The kernel representation foregrounds the relationship between causal theories and statistical experiments, and we will exploit the latter to justify the following claim: *Causal Bayesian Networks are a subset of causal theories*.

In fact, this holds for the following generalisations of causal graphical models:

- Arbitrary sets of CBNs (e.g. ADMGs)
- 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 $\{\mathsf{X}^i | i \in \mathbb{N}\}$ such that $\mathsf{X}^i : E \to X^i$. We denote an element $(x^0, \#, ..., \#, x^n) \in R$, $x^0, x^n \neq \#$ by the notation $do(\mathsf{X}^0 = x^0, \mathsf{X}^n = x^n)$ where occurrences of the distinguished element $*$ are ommitted. 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 **?**).

Strictly, we are considering labeled graphs $\mathcal{G}$ and sets $\{\mathsf{X}^i\}_{[n]}$ of random variables. That is, we have bijective functions $f : V \to [n]$ and $g : \{\mathsf{X}^i\}_{[n]} \to [n]$ and we adopt the convention that $f(i) := V^i$ and $g(i) := \mathsf{X}^i$. 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 $h : R \to \mathcal{P}([n])$ such that $h : (x^0, ..., x^n) \mapsto \{i | x^i \neq *\}$. That is, h picks out the indices that aren't suppressed in the $do(...)$ notation for elements of V. Define $\mathsf{X}^{i\prime} : R \to \{\#\} \cup X^i$ 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 $P_* : R \to \Delta(\mathcal{E})$ and $\{\mathsf{X}^i\}_{i\in[n]}$, 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 **?**)
2. For all $i \in h(r)$, $P_r F_{\mathsf{X}^i} = \delta_{\mathsf{X}^{i\prime}(r)}$
3. For all $i \notin h(r)$, $P_{r|\mathrm{Pa}_\mathcal{G}(\mathsf{X}^i)} F_{\mathsf{X}^i} = P_{\underline{\#}|\mathrm{Pa}_\mathcal{G}(\mathsf{X}^i)} F_{\mathsf{X}^i}$, $P_{\underline{\#}}$-almost surely

This definition differs slightly from that given in **?**; for example P_* is a map to $\Delta(\mathcal{E})$ rather than a set of labeled members of $\Delta(\mathcal{E})$, and we forumlate 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.

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

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 $P_*(\#) = \mu$. 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 $P_*(\#)$ positive definite, we have from this definition for any $r \in V$ the *truncated factorisation* property:

$$P_r F_{\mathbf{X}}(A) = \prod_{i\in h(r)} \delta_{\mathsf{X}^{i\prime}(r)}(\mathsf{X}^i(A)) \sum_{a\in A} \prod_{i\notin h(r)} P_{\#|\mathrm{Pa}_\mathcal{G}(\mathsf{X}^i)} F_{\mathsf{X}^i}(a; \{\mathsf{X}^i(a)\}) \tag{10}$$

As a consequence of the existence of conditional probability, given $\mathcal{G}$ and μ there exists a unique set

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

of interventional maps P_* compatible with both $\mathcal{G}$ and μ 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 $V \to \Delta(\mathcal{E})$. More generally, provided μ is compatible with $\mathcal{G}$ we have that **??** 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 $P_{\#|\mathrm{Pa}_{\mathcal{G}}(\mathsf{V}^i)}$ may be variously intended to mean a particular element of the class of conditional probabilities, any element or every element in the class class (**?**); condition 3 will have different implications for these various interpretations.

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_*

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 $\mathcal{T}^{\mathcal{G}} := \{(\mu, \kappa) | \mu \in \mathcal{H}, \kappa \in \mathcal{G}(\mu)\}$. 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 $\mathrm{Th} : \mathcal{G} \mapsto \mathcal{T}^{\mathcal{G}}$ 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 Th is injective)**.** *For DAGs $\mathcal{G}$, $\mathcal{G}'$ on the same set of RV's $\{\mathsf{X}^i\}_{[n]}$, $\mathcal{G} \neq \mathcal{G}' \implies \mathcal{T}^{\mathcal{G}} \neq \mathcal{T}^{\mathcal{G}'}$.*

Proof. Sketch: $\mathcal{G}$ and $\mathcal{G}'$ must disagree on at least one parental set. Choose some μ such that $P_{\underline{\#}|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)} F_{\mathsf{X}^i} \neq P_{\underline{\#}|\mathrm{Pa}_{\mathcal{G}'}(\mathsf{X}^i)} F_{\mathsf{X}^i}$. Then take 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 $r \neq r'$. Then $P_r F_{\mathsf{X}^i} \neq P_{r'} F_{\mathsf{X}^i}$ so $P_r \neq P_{r'}$. □

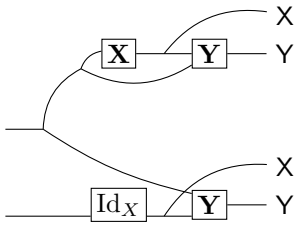
Each DAG $\mathcal{G}$ represents a causal theory $\mathcal{T}^{\mathcal{G}}$. For every causal theory $\mathcal{T}$, either it is not represented by any graph or there is a unique graph $\mathcal{G}$ such that $\mathcal{T} = \mathcal{T}^{\mathcal{G}}$. It is in this sense that we claim Causal Bayesian 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. **?** 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, 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 **?**)
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 $\mathcal{G} := X \twoheadrightarrow Y$, 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



(11)

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. (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 acutally trivial in this case; it is nontrivial where the graph is not fully connected). 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.

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

5 Potential outcomes models

We will follow **?** in our development of potential outcomes models, acknowledging that there are a wider variety of approaches to modelling potential outcomess than Rubin's version potential outcomes alone. Given an underlying state space Θ, Rubin posits a number of Markov kernels and composes them to define a statistical experiment $H : \Theta \to \Delta(\mathcal{X} \otimes \mathcal{Y} \otimes \mathcal{W})$. We eschew the "conditional probability" notation Rubin uses as it masks the distinction between a Markov kernel that is the disintegration of a previously defined joint probability distribution and a Markov kernel that we are simply defining as such, and will not necessarily be a disintegration of any probability distribution.

Notationally, we will refer to the symbol W_i as the i-th treatment assignment ($i \in \{0, .., n\}$), $\mathsf{Y}_i(0)$ $\mathsf{Y}_i(1)$ as the i-th potential outcomes, Y_i as the i-th obseved outcome and X_i as the i-th "vector of background facts". W refers to the composite of all W_is and similarly for other symbols. We avoid strictly defining what these symbols represent for now, as the construction of H allows us to be more explicit about what exactly these symbols represent. Suppose the vector $[\mathsf{Y}_0, ..., \mathsf{Y}_n]$ takes values in Y and similarly for other symbols.

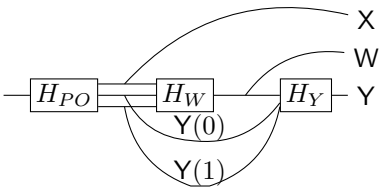
In particular, Rubin supplies:

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

Note that this construction does not permit the usual assumption of consistency ($W_i = w \implies \mathsf{Y}_i = \mathsf{Y}_i(w)$) because Markov kernels can at best give almost sure equality.

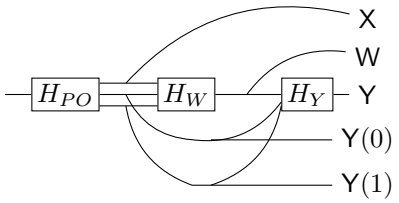
We then define the experiment H by

$H :=$



(12)

Where we have labeled the wires "carrying" $\mathsf{Y}(0)$ and $\mathsf{Y}(1)$ for clarity. Addtionally, we could draw an alternative diagram where each wire was copied n times to reflect the unit level variables. We can define an extended probability space H^* on which potential outcomes are also random variables:



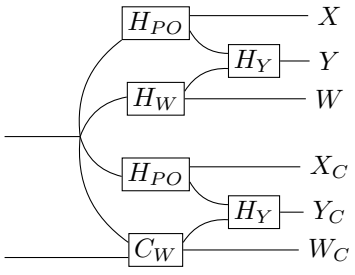
$H^* :=$ (13)

We are interested here in defining a "general type" of potential outcomes problem rather than investigate particular assumptions that may permit inference. Without a formal guide as to how to do this, we will postulate that a potential outcomes model is, in general, three Markov kernels $\langle H_{PO}, H_W, H_Y \rangle$ where $H_{PO} : \Theta \to \Delta(\mathcal{X} \otimes \mathcal{Y}^m)$, $H_W : \Theta \to \Delta(\{0, .., m-1\}^n)$ and $H_Y : \{0, .., m-1\}^n \times Y^m \to \Delta(\mathcal{Y})$ which is a "selection function" in the sense defined above. We adopt the alternative signature for H_W as it seems reasonable to suppose that the details of this kernel aren't always known *a priori*. Note that in general multiple potential outcomes models will yield the same statistical experiment H, though we postulate that in general different models will yield different H^*.

5.1 Can we consider potential outcomes models to be causal theories?

A potential outcomes model is a statistical experiment. Therefore, given a tuple $\langle \Theta, E, D, H, u \rangle$ where H follows from a potential outcomes model $\langle H_{PO}, H_W, H_Y \rangle$ and $u : F \to \mathbb{R}$ is a utility function, we have an ill-posed causal problem. A potential outcomes model is not literally a causal model, but we might ask if it is possible that our potential outcomes model H is a "causal theory in disguise"; Is there a natural map from potential outcomes models to causal theories?

We will propose, somewhat weakly, that given a well-specified potential outcomes model, decisions correspond to modifications of H_W. This supposition generalises the approach to policy modelling found in **?** As there is no general way to identify an arbitrary set of decisions D with different assignment functions H_W, we offer (weakly) that the answer to the question in the paragraph above is "no". However, given knowledge of the "decision-influenced treatment assignment" $C_W : \Theta \times D \to \{0, 1\}^n$, we *can* define a causal theory via the four elements $\langle H_{PO}, H_W, H_Y, C_W \rangle$. We've supposed here there are n "observational" units and n "consequence" units, a restriction that simplifies the notation and is fairly easy to lift. Concretely, the causal theory is:



$T :=$ (14)

Where X_C, W_C, Y_C are the "consequence" analogues of observational variables X, W and Y. To simplify the diagram, we have merged the wires for $\mathsf{Y}(0)$ and $\mathsf{Y}(1)$ and ommitted the labels; the potential outcomes are carried by the wire from H_{PO} to H_Y. Without detailed justification, we will note that this construction is unlikely to be appropriate if H_{PO} does not define an exchangeable sequence of potential outcomes, an assumption that we have made as a result of following **?**.

We can consider two cases where Eq. **??** appears to be appropriate.

First, suppose a potential outcomes model $\langle H_{PO}, H_W, H_Y \rangle$ is used in the evaluation of a public program, and it is intended to inform a choice between decisions $d = 0$: cut funding or $d = 1$: maintain funding. Suppose we also have C_W such that

- $d = 1$ leaves the assignment function unchanged; $C_W(\theta, 1; A) = H_W(\theta; A)$ for all θ, A

- d_0 means no-one receives treatment; $C_W(\theta, 0; A) = \delta_0(A)$ for all θ, A.

Supposing $Y = [0, 1]$ and positing a utility function $u := \pi_Y$, we can compare the utilities of decisions 0 and 1 for state θ by $Tu(\theta, 1) - Tu(\theta, 0)$ (in more familiar notation, $\mathbb{E}_{T(\theta,1;\cdot)}[u] - \mathbb{E}_{T(\theta,0;\cdot)}[u]$). By construction, if we let H^* be the "expanded" version of H above, $Tu(\theta, 1) - Tu(\theta, 0) = H^*_{\theta|\mathsf{W}}\pi_{\mathsf{Y}(1)}(1) - H^*_{\theta|\mathsf{W}}\pi_{\mathsf{Y}(0)}(1)$; this is because only the units for which $\mathsf{W} = 1$ have different outcomes under the different decisions. This quantity is known as the *effect of treatment on the treated* (ETT) (**?**).

ETT is common in the causal literature, but this is as far as I know the first example where it is formally derived as the difference between outcomes under different decisions, so I should probably do it properly

We can also consider the problem in medicine of evaluating the "effect of assigning treatment" vs "the effect of receiving treatment" (the former being known as *intention to treat* analysis). From **?**:

> 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.

As Shrier suggests, we can consider the potential outcomes $\mathsf{Y}(0)$ and $\mathsf{Y}(1)$ to represent the outcomes of an individual who is merely *assigned* a treatment or of an individual who is *actually given* a treatment. Note that in the former case, at least for a public health decision maker, one could reasonably suppose that the assignment function C_W was fully controlled by the decisions available - we can be absolutely certain, choosing $d = 1$, that the patient is assigned a treatment and likewise that they are not assigned for $d = 0$, for all θ. However, if the patient's chance to actually take the treatment may differ from experimental conditions - or even worse, if we are the patient and we *know* that we will take the treatment if we should decide to - then it is unclear how the potential outcomes model given would help to determine the expected consequences of a decision as the potential outcomes underlying the causal effect are now inappropriate. On the other hand, if we suppose that potential outcomes represent the outcomes of *taking* a treatment, then we can understand the difference between the three scenarios as differences in C_W - in the first case, we do not know C_W or H_W, but we do know that H_W factorises as $H_W = \curlyvee(I \otimes H_A)H'_W$ where $H_A : \Theta \to \Delta(A)$ is a treatment assignment function and $H'_W : \Theta \times A \to \Delta(W)$ is the treatment taking function. We also know $C_W = (I \otimes C_A)H'_W$ where $C_A : D \to \Delta(A)$ is a known assignment function. In the second case we do not know H_W or C_W nor do we necessarily have a relationship between the two, but we might posit some similarity. In the third case we know $C_W(\theta, d; \cdot)$ exactly – deciding to take the treatment means we definitely take the treatment and vise versa.

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

Theorem 6.1 (No causes in, no causes out (Bayes))**.** *If a causal theory $T : \Theta \times D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ and a prior $\xi \in \Delta(\Theta)$ are such that for all pairwise swaps $U_{dd'} : D \to \Delta(\mathcal{D})$, $(\xi \otimes U_{dd'})T = (\xi \otimes I)T$ 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:

$$\int_\Theta H_\theta J(\{d\}) C_\theta u(d) d\xi = \int_\Theta H_\theta J(\{d\}) U_{dd_0} C_\theta u(d) d\xi \tag{15}$$

$$= \int_\Theta H_\theta J(\{d\}) F_{\_d_0} C_\theta u(d) d\xi \tag{16}$$

$$\therefore \sum_{d \in D} \int_\Theta H_\theta J(\{d\}) F_{\_d_0} C_\theta(d; A) d\xi = \sum_{d \in D} \int_\Theta H_\theta J(\{d\}) C_\theta u(d) d\xi \tag{17}$$

$$= \int_\Theta \sum_{d \in D} H_\theta J(\{d\}) C_\theta u(d) d\xi \tag{18}$$

$$= \int_\Theta H_\theta J C_\theta u d\xi \tag{19}$$

$$= S_\xi(J) \tag{20}$$

$$= S_\xi(JF_{\_d_0}) \tag{21}$$

Where **??** follows from the fact that evaluation at d guarantees $U_{dd_0} C_\theta u(d) = F_{\_d_0} C_\theta u(d)$. □

Corollary 6.2. *If a causal theory T with a prior ξ 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 6.3 (No causes in, uniform strategy out (minimax))**.** *If a causal theory $T : \Theta \times D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ with finite D is such that for all pairwise swaps $U_{dd'} : D \to \Delta(\mathcal{D})$, $\theta \in \Theta$ there is some θ' 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 $D \to \Delta(\mathcal{D})$ 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 θ^0 where $S(J_d, \theta^0) = S(J_{d'}, \theta^0)$ for all $d, d' \in D$. Then $S(J, \theta^0) = S(J_U, \theta^0)$, contraticting our assumption that J achieved lower risk in the worst case. Suppose $S(J_U, \theta)$ is maximised in some state θ^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 $f : D \to D$ such that $f(A) \subset B$. But then there is some θ^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. □

Corollary 6.4. *If the risk of the uniform strategy is maximised in some state θ^* 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 vise 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

Then we ought to be indifferent between d_1 or d_2 whatever data we see.

No causes in, no causes out (NCINCO) implies that some common principles commonly applied to causal inference, in isolation, can only yield trivial theoreis. Without any notion of intervention, causal inference based solely on principles such as the invariance of conditionals **??**, a preference for low complexity consequences **?** or faithfulness **?** would yield triviality. As discussed in Section **??**, we also require assumptions on the effects of decisions to to get a causal theory from a potential outcomes model.

While decision biases are clearly a crucial element of useful causal theories, both major approaches to causality discussed here tend to favour a thorough treatment of assumptions that permit "inference" but a cursory treatment assumptions that yield decision biases. CBNs supply a generic notion of intervention that is unlikely to be a generally appropriate notion of the effects of a decision (see, for example, **?**). In addition, a CBN defines many more interventions than are every going to be at the disposal of a decision maker in practice – recall that we had to postulate a decision model for the CBN as well as the PO model in order to comapre the two. PO, on the other hand, appears to require a decision model to get off the ground at all. It appears to be the case that both approaches regard the details of "effects of decisions" as something that can be worked out later.

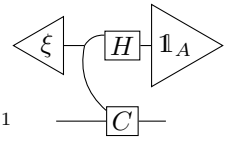
We postulate the concept of *modular extension* to formalise the notion of "working out the effects of decisions later". Informally, if a theory T permits a modular extension to T' then we can achieve the same result either by a) "doing inference" on T' directly or b) "doing inference" on T and then applying a decision model to yield T'. If T is not a useful theory, but it can be modularly extended to a large number of theories T' which we believe are useful, we may be best served by performing our analysis on T and saving the results for later. Proponents of both CBN and PO approaches appear to endorse the interpretation that the theories they produce represent "stable" knowledge of the real world.

We define this notion for theories equipped with a prior, and it is unclear if it can be extended to theories without a prior.

Definition 6.5 (Modular extension)**.** A theory $T : \Theta \times D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ equipped with a prior ξ permits modular extension to a theory $T' : \Theta \times D' \to \Delta(\mathcal{E} \otimes \mathcal{F})$ with the same prior ξ if there exists $M : D' \to \Delta(\mathcal{D})$ such that $(\xi \otimes \mathrm{Id}_D)T' = (\xi \otimes M)T$.

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

$$T_\xi|A := (\xi H(A))^{-1}$$

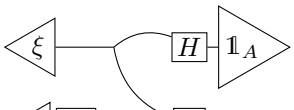


(22)

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

Theorem 6.6. *If T' is a modular extension of T under the shared prior ξ and module M, then for any strategy γ we have $T'_{\gamma,\xi}|A = \gamma M T_\xi|A$.*

Proof. By definition,



$$\gamma M T_\xi | A = (\xi H(A))^{-1} \tag{23}$$

$$= (\xi H(A))^{-1} \gamma(\xi \otimes M) T(\mathbb{1}_A \otimes \mathrm{Id}_F) \tag{24}$$

$$= (\xi H(A))^{-1} \gamma(\xi \otimes \mathrm{Id}_D) T'(\mathbb{1}_A \otimes \mathrm{Id}_F) \tag{25}$$

$$= T'_{\gamma,\xi} | A \tag{26}$$

□

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

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 θ_1 with "treatment efficacy" and θ_2 with "treatment susceptibility". Let $Y, W = \{0,1\}$ and suppose we have a potential outcomes model $H_{PO} : \Theta \to \Delta(\mathcal{Y}^2)$, $H_W : \Theta \to \Delta(\mathcal{W})$ and $H_Y : W \times Y^2 \to \Delta(\mathcal{Y})$. Furthermore, suppose we have $D = [0,1]^2$ and $C_W : \Theta_2 \times D \to \Delta(\mathcal{W})$ 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 $\Theta_2 \to \Delta(\mathcal{W})$. Then $\langle H_{PO}, H_W, H_Y, C_W \rangle$ defines a causal theory T and for $(d_1, d_2) \in D$, $A, B \in \mathcal{E}$:

$$T_\xi | A(d_1, d_2; B) = \frac{1}{\int_\Theta (H_{PO,\theta} \otimes H_{W,\theta}) H_Y(A)} \int_\Theta (H_{PO,\theta} \otimes H_{W,\theta}) H_Y(A) C_{W,\theta}(d_1, d_2; B) d\xi \tag{27}$$