# Story for AIStats

## 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 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 [Lewis, 1981], 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 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 $l : \Theta \times D \to \mathbb{R}$. 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 $\pi$ given state $\theta$ to be the expected loss of $\pi$ in state $\theta$. 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 $\Theta$; the coarsest measure rendering the evaluation maps measurable?

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 $\pi^*_{\mathrm{Ba}}$ such that the *Bayes risk* $R_\xi : \pi \mapsto \xi \curlyvee (H\pi \otimes \mathrm{Id}_\Theta) l$ is minimised:

$$\pi^*_{\mathrm{Ba}} \in \arg\min_{\pi \in \Pi} R_\xi(\pi) \tag{1}$$

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

$$\pi^*_{\mathrm{Mm}} \in \arg\min_{\pi\in\Pi} \max_{\theta\in\Theta} R(\theta,\pi) \tag{2}$$

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.

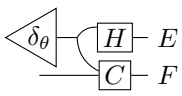
Our representation of statistical experiment is slightly different to, for example, Le Cam [1996], who introduces statistical experiments as an ordered collection of probability measures. Both representations do the same job, and the representation as a map makes for a clearer connection with causal 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 $\Theta \to \Delta(\mathcal{E})$ and $l$ a measurable function $E \to \mathbb{R}$. We leave implicit the set $\Pi$ 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 Ferguson [1967].

Observe that a statistical decision problem supplies a loss $l$ that tells us immediately how desirable a pair $(\theta, d) \in \Theta \times D$ is. 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 $\pi$), but we have no means of connecting decisions with consequences $F$. An obvious move is to introduce for each state of nature $\theta$ 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

$T_\theta :=$ 

(3)

Is sufficient to compute the causal risk. Thus we can replace $H$ and $C$ with the *causal theory kernel* $T : \Theta \times D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ 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 3, we can recover $H = T(\mathrm{Id}_E \otimes *_F)$ and $C = T(*_E \otimes \mathrm{Id}_F)$. Given $\alpha$ 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), $\alpha$ and $\beta$ 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 $\Theta$, $H$ 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 $\mathcal{T} = \{T_\theta | \theta \in \Theta\}$. Consider 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 $\Theta$ is an arbitrary set, for CSDPs we can regard $\Theta$ as a subset of kernels $D \to \Delta(\mathcal{E} \otimes \mathcal{F})$ that can be written as in Eq. 3 with the

this might be equivalent to the set of 1-combs

canonical kernel $\mathrm{Ev}_{\mathcal{T}}$. We call $\mathcal{T}$ the *causal theory set* and, as there is a bijection between theory kernels $T$ along with their domains $\Theta$ and theory sets $\mathcal{T}$, we typically refer to either as simply a *causal theory*. We say that $\alpha$ is a CSDP in kernel form and $\alpha'$ is a CSDP in set form.

## 3 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 more thorough definition see Pearl [2009]).

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 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 and unprimed random variables.

**Definition 3.1** (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 $\mathcal{G}$
2. For all $i \in h(r)$, $P_*(r)F_{\mathsf{X}^i} = \delta_{\mathsf{X}^{i\prime}(r)}F_{\mathsf{X}^i}$
3. For all $i \notin h(r)$, $P_*(r)_{|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)}F_{\mathsf{X}^i} = P_*(\underline{\#})_{|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)}F_{\mathsf{X}^i}$, $P_*(\underline{\#}; \cdot)$-almost surely

So far, this is a standard definition of a CBN; the extra additions are making explicit some implicit parts of the definition found in Pearl [2009].

Given a decision $y \in D$ (called a *do-intervention* in other treatments) and a distribution $\mu \in \Delta(\mathcal{E})$ that is *compatible* (Definition 3.2) with $\mathcal{G}$, $\mathcal{G}$ induces an *interventional* distribution $\mu^{\mathcal{G},y}$. The set of pairs $(\mu, y \mapsto \mu^{\mathcal{G},y})$ for $\mu$ compatible with $\mathcal{G}$ is a causal theory $\mathcal{T}_{\mathcal{G}}$.

In all following discussion, we assume the observed data represented by $\mathsf{X}$ is a sequence of independent and identically distributed random variables $\mathsf{X} = (\mathsf{X}_t)_{t\in T}$. We identify distributions over the sequence $\mathsf{X}$ with distributions over the initial observation $\mathsf{X}_0$ and subsequently drop the subscript.

The CBN convention is to denote an interventional distribution with $\mu(\cdot|do(\mathsf{X}^i = a))$. Here we associate every allowable set of $do$ statements with an element of the decision space $(D, \mathcal{D})$ equipped with random variables $\{\mathsf{D}^i\}_{i\in[N]}$ such that for $y \in D$, $\mu^y(\cdot) := P(\cdot|[do(\mathsf{X}^j = \mathsf{D}^i(y))]_{i\in N})$. The special element $*$ corresponds to a passive intervention which is denoted by the absence of a $do()$ statement in regular CBN notation.

**Definition 3.2** (Compatibility)**.** Given a DAG $\mathcal{G}$, $d$-separation is a ternary relation amongst sets of nodes the details for which we refer readers to Pearl [2009]. For a set of nodes $\{\mathsf{X}^i\}_{i\in[N]}$ we write $\mathsf{X}^i \perp_{\mathcal{G}} \mathsf{X}^j|\mathbf{X}$ to say $\mathsf{X}^i$ is d-separated in $\mathcal{G}$ from $\mathsf{X}^j$ by $\mathbf{X} \subset \{\mathsf{X}^i\}_{[N]}$.

Given a measurable space $(E, \mathcal{E})$, $\mu \in \Delta(\mathcal{E})$ and a set of random variables $\{\mathsf{X}^i\}_{i\in[N]}$ on $E$, $\mathsf{X}^i$ is independent of $\mathsf{X}^j$ conditional on $\mathbf{X}$ if $\mu_{|\mathbf{X}}\curlyvee(F_{\mathsf{X}^i} \otimes F_{\mathsf{X}^j}) = \mu_{|\mathbf{X}}F_{\mathsf{X}^i}\mu_{|\mathbf{X}}F_{\mathsf{X}^j}$, $\mu$-almost surely. This is written $\mathsf{X}^i \perp\!\!\!\perp_{\mu} \mathsf{X}^j|\mathbf{X}$.

$\mu$ is compatible with $\mathcal{G}$ if $\mathsf{X}^i \perp_{\mathcal{G}} \mathsf{X}^j|\mathbf{X} \implies \mathsf{X}^i \perp\!\!\!\perp_\mu \mathsf{X}^j|\mathbf{X}$

$\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)$ are the parents of $\mathsf{X}^i$ with respect to the graph $\mathcal{G}$ and $\mu_{|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)}$ is the conditional probability with respect to $\mu$ and the $\sigma$-algebra generated by the set $\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)$. Recall that $\mu \curlyvee (\otimes_{i \notin S(y)} F_{\mathsf{X}^i})$ is the joint distribution of $\{\mathsf{X}^i | i \in S(y)\}$.

To establish that the map $\kappa^{\mathcal{G},\mu} : D \to \Delta(\mathcal{X})$ given by $y \mapsto \mu^{\mathcal{G},y}$ is a consequence map, we must shown that it is measurable with respect to the $\sigma$-algebra generated by the set of variables $\mathsf{D}^i$; this is shown by Theorem **??** provided in Appendix **??**. Defining $\mathcal{H}_{\mathcal{G}} \subset \Delta(\mathcal{X})$ to be the set of distributions compatible with $\mathcal{G}$, the set of pairs $\{(\mu, \kappa^\mu) | \mu \in \mathcal{H}_{\mathcal{G}}\}$ is the causal theory $\mathcal{T}_{\mathcal{G}}$.

**Extending the theory induced by a CBN** The causal theory $T_{\mathcal{G}}$ defined above associates a consequence with every probability distribution compatible with $\mathcal{G}$ but not every probability distribution in $\Delta(\mathcal{X})$. It is arguably not reasonable to assume *a priori* that the conditional independences implied by $\mathcal{G}$ hold in the observed data. We might therefore regard the theory $\mathcal{T}_{\mathcal{G}}$ to be incomplete, and seek some extension of the theory for distributions not in $\mathcal{H}_{\mathcal{G}}$.

**Example 3.3** (Extension of a CBN)**.** Consider the graph $\mathcal{G} = C \longrightarrow A \longrightarrow B$ , which implies a single conditional independence: $\mathsf{C} \perp\!\!\!\perp \mathsf{B}|\mathsf{A}$.

Suppose the three associated random variables $\mathsf{A}$, $\mathsf{B}$ and $\mathsf{C}$ each take values in $\{0, 1\}$ and suppose (unrealistically) we know all $\mu$ in the set of possible joint distributions $\mathcal{H}$ share the marginal distribution $\mu F_{\mathsf{B}} := \zeta$ and the conditional distribution $\mu_{|\{\mathsf{A}\}} F_{\mathsf{B}} = \iota$ and $\mathsf{C}$ is "almost" independent of $\mathsf{B}$ given $\mathsf{A}$:

$$\max_{x \in \{0,1\}^3, y \in \{0,1\}} \left| \mu_{|\{\mathsf{A},\mathsf{C}\}} F_{\mathsf{B}}(x; \{y\}) - \iota(x; \{y\}) \right| < \epsilon \tag{4}$$

Suppose that only interventions on $\mathsf{A}$ are possible and the problem supplies a generalised utility such that, overloading $\mathsf{B}$, $U(\xi) = \mathbb{E}_\xi[\mathsf{B}]$. For convenience, we restrict our attention to the subset of decisions $D' = \{y | \mathsf{D}_{\mathsf{B}}(y) = \mathsf{D}_{\mathsf{C}}(y) = *\}$ and consequence maps marginalised over $\mathsf{A}$ and $\mathsf{C}$. Define $\kappa^{\mathcal{G}}$ by

$$\kappa^{\mathcal{G}}(y; Z) := \begin{cases} \iota(\mathsf{D}_A(y); Z) & \mathsf{D}_{\mathsf{A}}(y) \neq * \\ \zeta(Z) & \mathsf{D}_{\mathsf{A}}(y) = * \end{cases} \tag{5}$$

It can be verified that the causal theory $\mathcal{T}_{\mathcal{G}}$ induced by $\mathcal{G}$ and the set of compatible distributions $\mathcal{H}_{\mathcal{G}} \subset \mathcal{H}$ is the set of pairs $\{(\nu, \kappa^{\mathcal{G}}) | \nu \in \mathcal{H}_{\mathcal{G}}\}$.

Consider two options for extending this to distributions $\nu \in \mathcal{H}$ but not in $\mathcal{H}_{\mathcal{G}}$, noting that one could imagine many possibilities: $\mathcal{T}^{\subset}_{\mathcal{G}}$ is the union of causal theories given by all graphs $\mathcal{G}'$ on $\{A, B, C\}$ such that $\mathcal{G} \subset \mathcal{G}'$ (in this case, just $\mathcal{G}$ and $C \longrightarrow A \longrightarrow B$ ), and $\mathcal{T}^{\circ}_{\mathcal{G}}$ is the union of causal theories given by the all DAGs on the set of nodes $\{\mathsf{A}, \mathsf{B}, \mathsf{C}\}$.

The theory $\mathcal{T}^{\subset}_{\mathcal{G}}$ is given by $\mathcal{T}_{\mathcal{G}} \cup \{(\nu, \eta^\nu) | \nu \in \mathcal{H} \setminus \mathcal{H}_{\mathcal{G}}\}$ where

$$\eta^\nu := \begin{cases} (y; Z) \mapsto \sum_{c \in \{0,1\}} \nu F_{\mathsf{C}}(\{c\}) \nu_{|\{\mathsf{A},\mathsf{C}\}} F_{\mathsf{B}}(\mathsf{D}_A(y), c; Z) & \mathsf{D}_A(y) \neq * \\ \zeta(Z) & \mathsf{D}_A(y) = * \end{cases} \tag{6}$$

$\mathcal{T}^{\circ}_{\mathcal{G}}$ is the set of states associated with three types of graph: those featuring no arrow $A \not\longrightarrow B$ , those featuring $A \longrightarrow B$ but not $C \longrightarrow B$ and $C \longrightarrow A$ and the graph $C \longrightarrow A \longrightarrow B$ . These possibilities yield $\mathcal{T}^{\circ}_{\mathcal{G}} = \mathcal{T}^{\subset}_{\mathcal{G}} \cup \{(\nu, y \mapsto \zeta) | \nu \in \mathcal{H} \setminus \mathcal{H}_{\mathcal{G}}\}$.

By 4, $|\eta(x; \{y\}) - \iota(x; \{y\})| < \epsilon$ for all $x \in A \cup \{*\}$ and $y \in B$ and therefore for $J \in \mathcal{J}$, $|U(\mu J \curlyvee (I_{(D)} \otimes \eta)) - U(\mu J \curlyvee (I_{(D)} \otimes \iota))| < \epsilon$. Therefore a small $\epsilon$ ensures $\mathcal{T}^{\subset}_{\mathcal{G}}$ yields a risk set "close" to the risk given by $\mathcal{T}_{\mathcal{G}}$ for any $J$. On the other hand, $|\iota(x; \{y\}) - \zeta(\{y\})|$ is independent of $\epsilon$, so $\mathcal{T}^{\circ}_{\mathcal{G}}$ yields a risk set that contains points that do not converge to the risk set induced by $\mathcal{T}_{\mathcal{G}}$ with small $\epsilon$.

Extensions of the "base theory" $\mathcal{T}_{\mathcal{G}}$ can yield very different risk sets even when the departure from compatibility is slight and we limit those extensions to being based on CBNs. This example is

complementary to results indicating that with unknown variable ordering (which may be regarded as analogous to $\mathcal{T}^{\circ}_{\mathcal{G}}$) or with unmeasured confounders it is not possible to construct a test that uniformly converges to the true graph equivalence class [Robins et al., 2003, Zhang and Spirtes, 2003]; our example shows that some misses may be benign and others may not. We will finally note that the more general theory $\mathcal{T}^{\circ}_{\mathcal{G}}$ still has a nontrivial risk set, and hence (potentially) nontrivial implications for decision making. We think that the investigation of risk sets for "extended theories" discussed here or graph learning algorithms considered in the CBN literature presents many interesting questions.

## References


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

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.

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.

James M. Robins, Richard Scheines, Peter Spirtes, and Larry Wasserman. Uniform consistency in causal inference. *Biometrika*, 90(3):491–515, September 2003. ISSN 0006-3444. doi: 10.1093/biomet/90.3.491. URL https://academic.oup.com/biomet/article/90/3/491/231406.

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

Jiji Zhang and Peter Spirtes. Strong Faithfulness and Uniform Consistency in Causal Inference. In *Proceedings of the Nineteenth Conference on Uncertainty in Artificial Intelligence*, UAI'03, pages 632–639, San Francisco, CA, USA, 2003. Morgan Kaufmann Publishers Inc. ISBN 978-0-12-705664-7. URL http://dl.acm.org/citation.cfm?id=2100584.2100661. event-place: Acapulco, Mexico.