Causal Statistical Decision Problems

Anonymous Author(s)
Affiliation
Address
`email`

Abstract

We develop the notion of a causal statistical decision problem as an extension of the statistical decision theory of Wald. Suppose we have a dataset and some set of available decisions. Assume we know what state we would like the world to occupy, but we are uncertain about how our decisions affect the state of the world. We introduce the notion of *consequences* that relate decisions to states of the world, and *causal theories* that relate observations to consequences. A strength of this perspective is that it is not motivated by any notion of a "true cause" or "causal effect". We connect causal statistical decision problems to statistical decision problems and show that two leading approaches to causality - Causal Bayesian Networks and Potential Outcomes - have natural representations as causal theories. We argue that the causal theory associated with a CBN may be considered incomplete and discuss how different extensions can lead to very different properties.

1 Introduction

The decision theoretic approach to statistics casts statistical problems in terms of learning to output decisions that minimise a loss rather than learning true properties of a data generating distribution. Statistical decision theory plays a role of fundamental importance in modern machine learning; loss functions underpin the development of algorithms, and the analysis of losses is critical to the theoretical treatment of learning algorithms.

It is widely accepted that problems of causal inference are different to statistical problems. Causal problems are held to demand causal knowledge that is not in the vocabulary of statistical problems [Pearl, 2009, Cartwright, 1994]. There are two leading approaches to formalising "causal knowledge" and posing data-driven causal problems: one based on Causal Bayesian Networks and the other on Potential Outcomes.

Causal Bayesian Networks (CBNs) posit that there are causal relationships among a set of random variables that can be encoded by a directed acyclic graph (DAG). An investigator with access to the true graph and a joint probability distribution over all the variables present in that graph can calculate a wide variety of causal effects, and partial access to these objects will enable to partial knowledge of causal effects. A causal effect in this framework is is tied to the intuitive notion of "the result of intervening to set particular variables to particular values".

Potential Outcomes (PO) posits a large joint distribution over observed variables X, Y and partly unobserved "potential outcome" variables X_0, Y_1 and so forth. A potential outcome variable Y_i is interpreted as "the value of Y that would be observed if the action identified by i were taken". Under some conditions, an investigator with access to a joint distribution over observed variables may be able to infer certain properties of the distribution over potential outcome variables such as $\mathbb{E}[\mathsf{X}_i]$.

Queries in the CBN framework may be concerned with identification of causal effects given a graph and a probability distribution [Tian and Pearl, 2002], or with the determination of the true causal graph given just a probability distribution [Spirtes et al., 2000]. Queries in the PO framework usually

Submitted to 33rd Conference on Neural Information Processing Systems (NeurIPS 2019). Do not distribute.

concern identification of properties of the distribution of potential outcome variables known as *treatment effects* given a dataset and certain assumptions about this distribution [Rubin, 2005, Robins and Richardson, 2010]. In both cases, these queries fit the paradigm of "determining true properties of nature" rather than "learning to output a decision that minimises a loss".

The first contribution of this paper is the notion of a *causal statistical decision problem* (CSDP) that proceeds from a natural extension of an ordinary statistical decision problem (SDP) introduced by [Wald, 1950]. We suppose that, in contrast to an ordinary SDP where we have known preferences over (decision, state of nature) pairs, we know only our preferences over the *outcomes* of decisions, which we represent with a utility function. Uncertainty over the consequences of decisions is represented by a *causal theory* that connects observed data with *consequence maps*.

We show by a reduction that results concerning standard SDPs are also true of (at least) a subset of CSDPs. We also show that both Causal Bayesian Networks and joint distributions over potential outcomes have a natural representation as causal theories. Together these results show, for example, that the class of Bayes decision functions is a complete class for CSDPs based on Causal Bayesian Networks provided certain conditions on the utility and size of the available set of decisions are met.

The notion of a causal theory presented here can naturally represent models cast in terms of CBNs or POs, but there are many causal theories that cannot easily be represented by either. We discuss a question motivated by this more general perspective: *given a CBN with observable predictions, what should be assumed when the data doesn't match these predictions?* We show that different answers to this question yield widely divergent conclusions.

A key strength of our perspective is the possibility of theoretical treatment of causal learning from a viewpoint that is agnostic about the nature of "causal knowledge". Causal knowledge is a tricky domain from philosophy to practice, and there are many proposals for causal assumptions that do not neatly fit in either the CBN or PO camps [Bongers et al., 2016, Dawid, 2010, Bengio et al., 2019]. The theory presented here is capable of posing questions such as "does a proposed causal learning method work?" without first requiring commitments on the nature of causal knowledge. Substantial progress in machine learning has been the result of developing generic principles and learning techniques that are relevant to many datasets from many domains and are less reliant on the judgement of domain experts. We believe this separation of concerns is crucial to the advancement of generic techniques of causal learning.

Our approach is similar to that of Dawid [2012], but where he takes a "bottom-up" approach of developing a decision theoretic answer to particular causal questions, our approach is "top-down", proceeding from a general account of a causal problem to the particular objects needed to answer it. It also shares similarities with Causal Decision Theory developed by Lewis [1981], though the connection with statistical decision theory is better understood at this point.

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 while σ-algebras are calligraphic capitals $\mathcal{X}$ and 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. Sets of probability measures or stochastic maps are script capitals: $\mathscr{H}$, $\mathscr{T}$, $\mathscr{J}$.

A measurable space $(E, \mathcal{E})$ is a set E and a σ-algebra $\mathcal{E} \subset \mathcal{P}(\mathcal{E})$ containing the measurable sets. A probability measure $\mu \in \Delta(\mathcal{E})$ is a nonnegative map $\mathcal{E} \to [0, 1]$ such that $\mu(\emptyset) = 0$, $\mu(E) = 1$ and for countable $\{E_i\} \in \mathcal{E}$, $\mu(\cup_i E_i) = \sum_i \mu(E_i)$. We assume all measurable spaces discussed are standard. That is, they are isomorphic to either a subset of $\mathbb{N}$ with the discrete σ-algebra, or $\mathbb{R}$ with the Borel σ-algebra.

Given two measureable spaces $(E, \mathcal{E})$ and $(F, \mathcal{F})$, a *Markov kernel* or *stochastic map* $K : E \to \Delta(\mathcal{F})$ is a map where $x \mapsto K(x; B)$ is $\mathcal{E}$-measurable for every $B \in \mathcal{F}$ and $B \mapsto K(x; B)$ is a probability measure on $(F, \mathcal{F})$ for every $x \in E$. Abusing notation somewhat, we will write the set of Markov kernels of type $E \to \Delta(\mathcal{F})$ as $\Delta(\mathcal{F})^D$.

If we have two random variables $\mathsf{X} : \_ \to X$ and $\mathsf{Y} : \_ \to Y$, the conditional probability $P(\mathsf{Y}|\mathsf{X})$ is a Markov kernel $X \to \Delta(\mathcal{Y})$. Formally, given $\mu \in \Delta(\mathcal{E})$ and a sub-σ-algebra $\mathcal{E}' \subset \mathcal{E}$, there is a Markov kernel $\mu_{|\mathcal{E}'} : E \to \Delta(\mathcal{E})$ such that for $A \in \mathcal{E}$ and $B \in \mathcal{E}'$, $\int_B \mu_{|\mathsf{E}'}(y; A) d\mu(y) = \mu(A \cap B)$.

$\mu_{|\mathcal{E}'}$ is a *conditional probability distribution* with respect to $\mathcal{E}'$. This result may not hold if $(E, \mathcal{E})$ is not a standard measureable space [Çinlar, 2011].

Given a set of random variables $\mathbf{X} = \{\mathsf{X}^i\}_{i\in[N]}$ with domain $(E, \mathcal{E})$, $\mu_{\mathbf{X}} : E \to \Delta(\mathcal{E})$ is a conditional probability distribution with respect to the σ-algebra generated by $\mathbf{X}$: $\sigma(\cup_{i\in[N]}\sigma(\mathcal{X}^i))$. We will use this subscript notation rather than the more common bar notation (e.g. $\mu(\cdot|\mathbf{X})$) to express conditional probability from here onwards.

Two Markov kernels $K : E \to \Delta(\mathcal{F})$ and $K' : E \to \Delta(\mathcal{F})$ are μ-almost surely equivalent given $\mu \in \Delta(\mathcal{E})$ if for all $A \in \mathcal{E}, B \in \mathcal{F}$, $\int_A K(x; B)d\mu = \int_A K'(x; B)$.

Kernel products: Kernel products allow common operations to be written compactly. The notation here borrows heavily from Çinlar [2011] and Fong [2013]. More details can be found in Appendix A. For the following, assume $K : E \to \Delta(\mathcal{F})$, $L : F \to \Delta(\mathcal{G})$, and $M : G \to \Delta(\mathcal{H})$ are Markov kernels, μ is a probability measure on $(E, \mathcal{E})$.

The *kernel-kernel* product KL is a Markov kernel $E \to \Delta(\mathcal{G})$ such that $KL(x; B) := \int_F K(x; dy)L(y; B), \qquad x \in E, B \in \mathcal{G}$. Kernel-kernel products are associative: $(KL)M = K(LM)$.

The *measure-kernel* product of μ and K, μK is a probability measure on $(F, \mathcal{F})$ such that $\mu K(B) = \int_E \mu(dx)K(x; B), \qquad B \in \mathcal{F}$. Measure-kernel products are also associative: $(\mu K)L = \mu(KL)$.

Special kernels: $I_{(E)}$ is the identity kernel $E \to \Delta(\mathcal{E})$ defined by $x \mapsto \delta_x$. It has the properties $\mu I_{(E)} = \mu$, $KI_{(F)} = K$, $I_{(E)}K = K$.

Given some measurable function $g : E \to F$, the kernel $F_g : E \to \Delta(\mathcal{F})$ is defined by $x \mapsto \delta_{g(x)}$. It is easy to check that $F_g F_g = F_g$. For $\mu \in \Delta(\mathcal{E})$, $\mu F_g(A) = \mu(g^{-1}(A))$. This notation allows us to consistently represent a marginal distribution μF_{X} and a marginal kernel κF_{X}.

Given $\mu \in \Delta(\mathcal{E}, \mu^{\curlyvee}(I_{(E)} \otimes K)$ is a distribution in $\Delta(\mathcal{E} \otimes \mathcal{F})$ given by

$$\mu^{\curlyvee}(I_{(E)} \otimes K)(A \times B) = \int_A K(x; B)d\mu(x) \qquad \forall A \in \mathcal{E}, B \in \mathcal{F} \tag{1}$$

The symbol $\curlyvee$ is read "splitter".

3 Causal Statistical Decision Problems

	SDPs	CSDPs
State of the world	$\mathcal{H}$, hypothesis class	$\mathcal{T}$, causal theory
Observations	X	X
Decisions	D	D
Known preferences	ℓ, loss	U, generalised utility
Derived preferences	ℓ, loss	L, causal loss

Table 1: Comparison of SDPs and CSDPs

We develop causal statistical decision problems (CSDPs) inspired by statistical decision problems (SDPs) of Wald [1950]. CSDPs differ from SDPs in that our preferences (i.e. utility or loss) are known less directly in former case. We show that every SDP can be represented by a CSDP and that the converse is sometimes but not always possible. We show that an analogoue of the fundamental *complete class theorem* of SDPs applies to the class of CSDPs that can be represented by SDPs, but whether such a theorem applies more generally is an open question.

Following [Ferguson, 1967], we consider SDPs and CSDPs to represent normal form two person games. At the most abstract level the games represent the options and possible payoffs available to the decision maker, and this representation allows us to compare the two types of problem. In their more detailed versions, CSDPs and SDPs differ in their representation of the state of the world and in the type of function that represents preferences. These differences are summarised in Table 1.

Definition 3.1 (Normal form two person game)**.** A normal form game is a triple $\langle \mathcal{S}, A, L\rangle$ where $\mathcal{S}$ and A are arbitrary sets and $L : \mathcal{S} \times A \to [0, \infty)$ is a loss function.

The set $\mathcal{S}$ is a set of possible states that the environment may occupy and A is a set of actions the decision maker may take. The decision maker seeks an action in A that minimises the loss L. Generally there is no action that minimises the loss for all environment states. A minimax solution is an action that minimises the worst case loss: $a^*_{mm} = \arg\min_{a\in A}[\sup_{s\in\mathcal{S}} L(s,a)]$.

If the set $\mathcal{S}$ is equipped with a σ-algebra $\mathcal{S}$ and a probability measure $\xi \in \Delta(\mathcal{S})$ which we will call a "prior", a Bayes solution minimizes the expected risk with respect to ξ: $a^*_{ba} = \arg\min_{a\in A} \int_{\mathcal{S}} L(s,a)\xi(ds)$.

Definition 3.2 (Admissible Action)**.** Given a normal form two person game $\langle \mathcal{S}, A, L\rangle$, an action $a \in A$ is *strictly better* than $a' \in A$ iff $L(s,a) \leq L(s,a')$ for all $s \in \mathcal{S}$ and $L(s_0,a) < L(s_0,a')$ for some $s_0 \in \mathcal{S}$. If only the first holds, then a is as good as a'. An *admissible action* is an action $a \in A$ such that there is no action strictly better than A.

Definition 3.3 (Complete Class)**.** A class C of decisions is a *complete class* if for every $a \notin C$ there is some $a' \in C$ that is strictly better than a.

C is an *essentially complete* class if for every $a \notin C$ there is some $a' \in C$ that is as good as a.

Definition 3.4 (Reduction)**.** A normal form two person game $\alpha = \langle \mathcal{S}^\alpha, A, L^\alpha\rangle$ can be reduced to a different game sharing the same action set $\beta = \langle \mathcal{S}^\beta, A, L^\beta\rangle$ if there is a surjective function $g : \mathcal{S}^\alpha \to \mathcal{S}^\beta$ such that for every $a \in A$, $s \in \mathcal{S}^\alpha$, $L^\alpha(s,a) = L^\beta(g(s),a)$.

Because CSDPs and SDPs posit states of nature of different types, they cannot represent exactly the same game. Reduction is a notion that allows us to say tht the game represented by a CSDP and an SDP are "essentially" the same. Reduction preserves the important properties of admissibility and completeness as shown in Appendix B.

A statistical decision problem represents a normal form two-person game where the available actions are *decision functions* that output a decision given data, the states of the environment are associated with probability measures on some measurable space and we assume a loss expressing preferences over decisions and states is known.

Definition 3.5 (Statistical Decision Problem)**.** A statistical decision problem (SDP) is a triple $\langle \mathcal{H}, D, \ell\rangle$. $\mathcal{H} \subset \Delta(\mathcal{E})$ is a hypothesis class representing possible states of the environment, D is the set of available decisions, $\mathsf{X} : (E,\mathcal{E}) \to (X,\mathcal{X})$ is a random variable representing the information available for the statistician to make a decision and $\ell : \mathcal{H} \times D \to [0,\infty)$ is a loss function.

Denote by $\mathcal{J}$ the set of decision kernels $X \to \Delta(\mathcal{D})$. Recall that $\mu F_{\mathsf{X}}(A) = \mu(\mathsf{X}^{-1}(A))$. For $J \in \mathcal{J}$ and $\mu \in \mathcal{H}$, the risk $R : \mathcal{J} \times \mathcal{H} \to [0,\infty)$ is defined as $R(J,\mu) = \int_D \ell(\mu,y)\mu F_{\mathsf{X}} J(dy)$. The triple $\langle \mathcal{H}, \mathcal{J}, R\rangle$ forms a two player normal form game.

The loss function ℓ expresses preferences over (state, decision) pairs. However, it may be the case that our preferences don't naturally apply directly to such pairs. For a doctor deciding whether to prescribe a treatment to a patient, it is clear that this patient being healthy in the future is preferable to them being sick. This motivates the definition of a causal statistical decision problem, utilising a preferences defined over outcomes represented by a generalised utility rather than (state, decision) pairs. A generalised utility, as opposed to an ordinary utility, allows for a simple reduction from CSDPs to SDPs. In order to compute the loss associated with a decision a map from decisions to outcomes is required, which we term a *consequence*.

Definition 3.6 (Consequences)**.** Given a measurable outcome space $(F,\mathcal{F})$ and a measurable decision space $(D,\mathcal{D})$, a Markov kernel $\kappa : D \to \Delta(\mathcal{F})$ is a *consequence mapping*, or just a consequence for short.

Definition 3.7 (Causal state)**.** Given a consequence $\kappa : D \to \Delta(\mathcal{F})$, a measurable observation space $(E,\mathcal{E})$ and some distribution $\mu \in \Delta(\mathcal{E})$, the pair $\tau := (\kappa,\mu)$ is a *causal state* on D, E and F. We refer to κ as the consequence and μ as the observed state.

We allow the "observation" space E and the "outcome" space F to differ as it it may be desirable to avoid modelling consequences on variables that are observed but irrelevant to preferences (see Theorems B.6 and B.5). In practice these spaces often coincide.

Definition 3.8 (Causal Theory)**.** A causal theory $\mathcal{T}$ is a set of causal states sharing the same decision, observation and outcome spaces. We abuse notation to assign the "type signature" $\mathcal{T} : E \times D \rightarrowtriangle F$ for a causal theory with observed distributions in $\Delta(\mathcal{E})$ and consequences $D \to \Delta(\mathcal{F})$.

Definition 3.9 (Causal Statistical Decision Problem)**.** A causal statistical decision problem (CSDP) is a tuple $\langle(\mathcal{T},(E,\mathcal{E})\mathsf{X}),(D,\mathcal{D}),(U,(F,\mathcal{F}))\rangle$. $\mathcal{T}$ is a causal theory on D, E and F, D is the decision set, $\mathsf{X}:(E,\mathcal{E})\to(X,\mathcal{X})$ is a random variable representing the given information and $U:\Delta(\mathcal{F}\otimes\mathcal{D})\to\mathbb{R}$ is a generalised utility expressing preference over joint distributions of decisions and outcomes which we assume is bounded above.

From the generalised utility U we can define a loss $L:\mathcal{T}\times\Delta(\mathcal{D})\to[0,\infty]$ by

$$L((\kappa,\mu),\gamma):=\sup_{\gamma'\in\Delta(\mathcal{D})}U(\gamma'\curlyvee(I_{(D)}\otimes\kappa))-U(\gamma\curlyvee(I_{(D)}\otimes\kappa))\tag{2}$$

For $(\kappa,\mu)\in\mathcal{T}$ and $\gamma\in\Delta(\mathcal{D})$. This is well defined wherever U is bounded above. Note that L does not depend on the data generating distribution μ; henceforth we will suppress this argument and write $L(\kappa,\gamma):=L((\kappa,\mu),\gamma)$.

Given a decision function $J\in\mathcal{J}$ and $(\kappa,\mu)\in\mathcal{T}$, we define the risk $R:\mathcal{J}\times\mathcal{T}\to[0,\infty)$ by $R(J,\kappa,\mu):=L(\kappa,\mu F_{\mathsf{X}}J)$. The triple $\langle\mathcal{T},\mathcal{J},R\rangle$ is a normal form two person game.

If there exists some measurable $u:F\times D\to\mathbb{R}$ such that for all $\xi\in\Delta(\mathcal{F}\otimes\mathcal{D})$, $U(\xi)=\mathbb{E}_\xi[u]$ then we call U an ordinary utility. An ordinary induces a loss $L(\kappa,\gamma)=\mathbb{E}_\gamma[l^\kappa]$ where $l^\kappa:D\to[0,\infty)$ is defined by

$$l^\kappa(d):=\sup_{\gamma'\in\Delta(\mathcal{D})}\mathbb{E}_{\gamma'\curlyvee(I_{(D)}\otimes\kappa)}[u]-\mathbb{E}_{\kappa(d;\cdot)}[u(\cdot,d)]\tag{3}$$

Reduction from a CSDP to an ordinary SDP is sufficient to import results from statistical decision theory such as Theorem 3.10. The complete class theorem along with Corollary 3.14 show that, at least for CSDPs with countable decision sets, finite causal theories and ordinary utilities, any admissible decision rule is a Bayes rule given *some* prior on $\mathcal{T}$. Stronger versions of Theorem 3.10 exist for SDPs, and stronger versions are likely to exist for CSDPs as well.

Theorem 3.10 (Complete class theorem (CSDP))**.** *Given an CSDP $\alpha:=\langle(\mathcal{T},E),D,\mathsf{X},U\rangle$ with risk R, if there is a reduction to an SDP $\beta:=\langle(\mathcal{H},F),D,\mathsf{Y},\ell\rangle$ with risk R' such that $|\mathcal{H}|<\infty$ and $\inf_{J\in\mathcal{J},\mu\in\mathcal{H}}R'(J,\mu)<-\infty$, then the set of all Bayes decision functions is a complete class for β and the set of all admissible Bayes decision functions is a minimal complete class for β.*

Proof. This follows from Lemmas B.2 and B.3. See appendix B. $\square$

Any statistical decision problem can be reduced to a CSDP featuring a causal theory where decisions have no effect.

Theorem 3.11. *Every SDP $\langle(\mathcal{H},E,\mathsf{X}),D,\ell\rangle$ can be reduced to a CSDP.*

Proof. Choose a theory that matches every probability measure $\mu\in\Delta(\mathcal{E})$ with a consequence map that itself always yields μ. Then construct a generalised utility U that induces an identical risk. See Appendix B. $\square$

A CSDP cannot, in general, be reduced to a statistical decision problem - for example, if we choose the utililty to be the variance of some random variable we may be able to achieve higher utility through a randomised decision than any nonrandomised decision under conditions where a regular SDP cannot have this property (see Example B.9 in Appendix B). It is an open question whether this reduction is generally possible if the problem features a ordinary utility.

Theorem 3.12 reduces a CSDP to a SDP by associating each pair (μ,κ) in the causal theory with a distribution over $E\times F\times D$. This may be possible by finding a distribution on the product space for which μ is a marginal probability and κ is a conditional distribution, but this is not necessary.

Theorem 3.12. *Given a CSDP $\beta=\langle(\mathcal{T},E,\mathsf{X}),D,(U,F)\rangle$ where U is an ordinary generalised utility, let $\mathcal{K}=\{\kappa|(\kappa,\mu)\in\mathcal{T}\}$ be the set of consequences. β is reducible to a statistical decision problem on the measurable space $(E\times F\times D,\mathcal{E}\otimes\mathcal{F}\otimes\mathcal{D})$ if there is some surjective map $m:\Delta(\mathcal{F}\otimes\mathcal{D})\to\mathcal{K}$.*

Proof. We construct the map h based on the map m and show that, given an ordinary utility U, it is possible to construct a loss ℓ such that the resulting SDP features the same risk assignments as the original CSDP. See Appendix B. $\square$

Corollary 3.13. *If the cardinality of $\Delta(\mathcal{F} \otimes \mathcal{D})$ is at least as large as the cardinality of the set of Markov kernels $D \to \Delta(\mathcal{F})$ then an CSDP with an ordinary utility can always be reduced to a SDP.*

A major open question, then, is if $(E, \mathcal{E})$ and $(D, \mathcal{D})$ are standard measurable spaces, the conditions for Corollary 3.13 hold in general. Corollary 3.14 shows that the reduction can be made in general if D is a denumerable set.

Corollary 3.14. *A CSDP $\langle(\mathcal{T}, (E, \mathcal{E}), \mathsf{X}), (D, \mathcal{D}), (U, (F, \mathcal{F}))\rangle$ where D is a denumerable set and U is an ordinary generalised utility can be reduced to a statistical decision problem.*

Proof. Take some probability measure $\pi \in \Delta(\mathcal{D})$ such that $\pi(\{y\}) > 0$ for all $y \in D$. Such a π exists by the denumerability of $\mathcal{D}$. The map $m : \Delta(\mathcal{F} \otimes \mathcal{D}) \to \mathcal{K}$ given by $m(\xi)(y; A) := \frac{\xi(A \times \{y\})}{\pi(\{y\})}$ is surjective. The result follows from Theorem 3.12. □

4 Causal Bayesian Networks

A Causal Bayesian Network (CBN) is a directed acyclic graph (DAG) $\mathcal{G}$ containing a set of nodes $\{\mathsf{X}^i\}_{i\in[N]}$ which we identify with random variables on some space $(E, \mathcal{E})$. 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 4.1) 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 μ compatible with $\mathcal{G}$ is a causal theory $\mathcal{T}_{\mathcal{G}}$.

In all following discussion, we assume the observed data represented by 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 X with distributions over the initial observation 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 4.1 (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 X^i is d-separated in $\mathcal{G}$ from 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, X^i is independent of X^j conditional on $\mathbf{X}$ if $\mu_{|\mathbf{X}}\curlyvee(F_{\mathsf{X}^i} \otimes F_{\mathsf{X}^j}) = \mu_{|\mathbf{X}} F_{\mathsf{X}^i} \mu_{|\mathbf{X}} F_{\mathsf{X}^j}$, μ-almost surely. This is written $\mathsf{X}^i \perp\!\!\!\perp_\mu \mathsf{X}^j|\mathbf{X}$.

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

Definition 4.2 (Causal Bayesian Network)**.**

A CBN has a graph $\mathcal{G}$ with edges $\{V^i\}_{[N]}$, random variables $\{\mathsf{X}^i\}_{[N]}$ and decision variables $\{\mathsf{D}^i\}_{[N]}$ which are all associated with one another in the obvious way. It would be nice to have a simple way of expressing this bundle of things and the corresponding associations

Consider a directed acyclic graph $\mathcal{G}$ with nodes $\mathbf{X} = \{\mathsf{X}^i | i \in [N]\}$, a measurable space $(E, \mathcal{E})$ and a set of random variables $\mathsf{X}^i : E \to X^i$ and $X = \times_{i\in[N]} X^i$ along with decision space $(D, \mathcal{D})$ and random variables $\{\mathsf{D}^i\}_{i\in[N]}$ where $\mathsf{D}^i : D \to X^i \cup \{*\}$.

Given any $y \in D$ let $S(y) \subset [N]$ be the set of all indices i such that $\mathsf{D}^i(y) \neq *$. Let $\mathcal{H}_{\mathcal{G}} \subset \Delta(\mathcal{X})$ be the set of distributions compatible with $\mathcal{G}$. Given arbitrary $\mu \in \mathcal{H}_{\mathcal{G}}$ and $y \in D$ the $\mathcal{G}, \mu, y$-interventional distribution denoted $\mu^{\mathcal{G},y}$ is given by the following three conditions:

1. $\mu^{\mathcal{G},y}$ is compatible with $\mathcal{G}$
2. For all $i \in S(y)$, $\mu^{\mathcal{G},y} F_{\mathsf{X}^i} = \delta_{\mathsf{D}^i(y)} F_{\mathsf{X}^i}$
3. For all $i \notin S(y)$, $\mu^{\mathcal{G},y}_{\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)} F_{\mathsf{X}^i} = \mu_{|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)} F_{\mathsf{X}^i}$, $\mu^{\mathcal{G},y}$-almost surely

$\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)$ are the parents of 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 μ and the σ-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 σ-algebra generated by the set of variables D^i; this is shown by Theorem C.1 provided in Appendix C. 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 4.3 (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 A, B and C each take values in $\{0, 1\}$ and suppose (unrealistically) we know all μ 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 C is "almost" independent of B given 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 A are possible and the problem supplies a generalised utility such that, overloading 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 A and 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: $\mathcal{T}_{\mathcal{G}}^{\subseteq}$ 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 $\begin{tikzcd} C \arrow[r] \arrow[rr, bend left] & A \arrow[r] & B \end{tikzcd}$), and $\mathcal{T}_{\mathcal{G}}^{\circ}$ is the union of causal theories given by the all DAGs on the set of nodes $\{\mathsf{A}, \mathsf{B}, \mathsf{C}\}$.

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

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

$\mathcal{T}_{\mathcal{G}}^{\circ}$ 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 $\begin{tikzcd} C \arrow[r] \arrow[rr, bend left] & A \arrow[r] & B \end{tikzcd}$. These possibilities yield $\mathcal{T}_{\mathcal{G}}^{\circ} = \mathcal{T}_{\mathcal{G}}^{\subseteq} \cup \{(\nu, y \mapsto \zeta)|\nu \in \mathcal{H} \setminus \mathcal{H}_{\mathcal{G}}\}$.

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 ϵ ensures $\mathcal{T}_{\mathcal{G}}^{\subseteq}$ 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 ϵ, so $\mathcal{T}_{\mathcal{G}}^{\circ}$ yields a risk set that contains points that do not converge to the risk set induced by $\mathcal{T}_{\mathcal{G}}$ with small ϵ.

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 $\mathcal{T}_{\mathcal{G}}^{\circ}$) 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 $\mathcal{T}_{\mathcal{G}}^{\circ}$ 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.

5 Potential Outcomes

Potential Outcomes is an alternative to the approach typified by Causal Bayesian Networks for formulating causal questions and hypotheses. Causal queries in the Potential Outcomes framework concern the distribution of random variables $\mathsf{X}_0, \mathsf{X}_1$ representing potential outcomes, or "the value X would have taken if action 0 or 1 were taken respectively" (Hernán and Robins [2018]). This is similar, but not the same, as the question answered by a consequence map which is "what is the distribution of X if I take actions 0 or 1?"

A natural connection between these informal notions of potential outcomes and consequence maps is given by the notion of consequence consistency. Let $\Delta(\mathcal{Y}_\circ)$ be the space of joint distributions over real and potential outcomes of X. A consequence map $\kappa : D \to \Delta(\mathcal{Y}_\circ)$ is consequence consistent if

$$(\delta_i \kappa)_{|\mathsf{X}_i} F_{\mathsf{X}}(w; A) = \delta_{\mathsf{X}_{i(w)}}(A) \tag{7}$$

Consequence consistency is similar to the consistency condition [Richardson and Robins, 2013], but the latter does not involve consequences.

A causal theory that is consequence consistent need not have any particular relationship between an "observed" distribution $\mu \in \Delta(\mathcal{Y}_\circ)$ and an associated consequence κ; one choice to make this connection is equality of the distributions of potential outcomes $\mu F_{\mathsf{X}_i} = \delta_i \kappa F_{\mathsf{X}_i}$, $i \in D$. Example D.1 in Appendix D shows that other choices may be preferred.

6 Equivalence of causal problems

Under what conditions could we consider a consequence consistent theory $\mathcal{T}^{cc}$ associated with some distribution over potential outcomes to be "equivalent" to some causal theory $\mathcal{T}^{\mathcal{G}}$ associated with a CBN $\mathcal{G}$ or vise versa?

The question of whether $\mathcal{T}^{\mathcal{G}}$ is consequence consistent with respect to some distribution over potential outcomes is easy to answer in the affirmative as consequence consistency is a trivial requirement if we choose potential outcomes $\mathsf{X}_y := \mathsf{X}$ for all $y \in D$.

The question of whether a consequence consistent theory $\mathcal{T}^{cc}$ can in general be represented by a Causal Bayesian Network is then also straightforwardly answered in the negative, as conditions 2 and 3 of Definition 4.2 are in general non-trivial (condition 1 is trivial given a fully connected DAG $\mathcal{G}$).

The trivial potential outcome $\mathsf{X}_y = \mathsf{X}$ clashes with the informal idea that a potential outcome represents the value X would have taken had action y been taken - we might expect, for example, if $\delta_y \kappa F_{\mathsf{X}} \neq \mu F_{\mathsf{X}}$ then X would at least sometimes take a different value if the action y is taken than if it is not.

We might tentatively propose a more extensive set of assumptions to characterise a "Potential Outcomes" theory, which we will write $\mathcal{T}^{po}$.

Definition 6.1 (Potential Outcomes Causal Theory)**.** A causal theory $\mathcal{T}^{po}$ is a "Potential Outcomes" theory with respect to random variable $\mathsf{X} : E \to X$ and potential outcome variable $\mathsf{X}_i : E \to X$, $i \in D$ if for every $(\mu, \kappa) \in \mathcal{T}$, κ is consequence consistent (Eq. 7) and

$$\mu F_{\mathsf{X}_i} = \delta_i \kappa F_{\mathsf{X}_i} \tag{8}$$

If we consider only joint distributions over potential outcomes, a PO causal theory associates a unique consequence with each distribution

Note that the condition of consistency [Richardson and Robins, 2013], which is a very standard condition in the Potential Outcomes literature, is:

$$\mu_{|\{\mathsf{X}_i, \mathsf{Z}\}} \mathsf{F}_{\mathsf{X}}(w; A) = \delta_{\mathsf{X}_{i(w)}}(A) \qquad w \in \mathsf{Z}^{-1}(i) \tag{9}$$

Where the random variable Z is a variable that is informally understood to be "intervenable" in a similar manner to intervention in Causal Bayesian Networks. A Potential Outcomes Causal Theory invokes a very general notion of Potential Outcomes where such intervenable variables may not exist, and so consistency may not be a sensible notion.

We can specify causal theories with a CBN $\mathcal{G}$ that are not potential outcomes causal theories. Consider the graph X (with a single node and no edges). By condition 2 of Definition 4.2, the consequences in $\mathcal{T}^{\mathcal{G}}$ will all yield X distributed as a delta function for certain decisions. However, in general $\mathcal{T}^{\mathcal{G}}$ will contain distributions on the observation space E for which no variable is distributed according to a delta function. $\mathcal{T}^{\mathcal{G}}$ therefore cannot be a Potential Outcomes Causal Theory. We will outline below how a Potential Outcomes theory is not, in general, a theory associated with any CBN $\mathcal{G}$.

Rather than demand that we can represent the same theory with a CBN and with PO, we might ask instead if a problem featuring a PO theory can in general be reduced to a problem featuring a CBN theory and vise versa. This is in keeping with our approach that a CSDP represents at a high level a two person game and the latter determines the decision-relevant aspects of the problem.

Definition 6.2 (Potential Outcomes CSDP)**.** A CSDP $\langle(\mathcal{T},(E,\mathcal{E}),\mathsf{X}),(D,\mathcal{D}),(U,(F,\mathcal{F})\rangle$ is a *Potential Outcomes CSDP* (POCSDP) if $E=F$, D is denumerable and there exists a set of potential outcome variables $\mathsf{X}_i : E \to X$, $i \in D$ with respect to which $\mathcal{T}$ is a Potential Outcomes causal theory.

Definition 6.3 (CBN CSDP)**.** A CSDP $\langle(\mathcal{T},(E,\mathcal{E}),\mathsf{X}),(D,\mathcal{D}),(U,(F,\mathcal{F})\rangle$ is a *Causal Bayesian Network CSDP* (CBNCSDP) with respect to some finite DAG $\mathcal{G}=(V,W)$ if $E=F$ and $\mathcal{T}$ is the theory induced by $\mathcal{G}$

(...and all the other stuff you need).

Theorem 6.4 shows that, supposing D is denumerable, every CSDP can be reduced to a PO CSDP. For denumerable D, then, it suffices to show that conditions 1-3 of Definition 4.2 are nontrivial. Take some CSDP $\alpha=\langle(\mathcal{T},E,\mathsf{X}),D,(U,E)\rangle$ and suppose there is no $(\kappa,\mu)\in\mathcal{T}$, $y\in D$, $z\in X$ such that $\delta_y\kappa F_{\mathsf{X}}(A)=\delta_z(A)$. Then it is straightforward to see that α cannot satisfy condition 3 of Definition 4.2. Suppose that there is no $(\kappa,\mu)\in\mathcal{T}$, $y\in D$ such that $\delta_y\kappa=\mu$; it is then straightforward that conditions 1 and 2 of Definition 4.2 cannot be simultaneously satisfied.

In both cases it is straightforward to posit generalised utilities such that α cannot be reduced.

Lifting condition 2 from the definition of a CBN yields CBNs with *generalized interventions*.

I *strongly suspect* this corresponds to the class of influence diagrams of [Dawid, 2010]

. Because conditions 1+2 are nontrivial, there exist POCSDPs that cannot be reduced to CSDPs based on CBNs with generalised interventions. Lifting conditions 2 and 3 yields a causal theory where we require only that the distributions given by every consequence κ are compatible with some DAG $\mathcal{G}$, which we will call an *independence-only CBN*

I *strongly suspect* this is closely related to the notion of Extended Conditional Independence of [Dawid, 2012]

. Condition 1 of Definition 4.2 can always be satisfied by choosing a graph $\mathcal{G}$ that is fully connected, so lifting conditions 2 and 3 is sufficient to ensure that every POCSDP can be reduced to a CSDP featuring an independence-only CBN, and in fact an independence-only CBN can represent every PO causal theory.

The single world intervention graphs of Richardson and Robins [2013] are DAGs that represent independences among distributions over potential outcome variables. They might be interpretable as POCSDPs.

The generalised versions of CBNs yield theories that generally associate multiple consequences with each given distribution. However a generalized CBN still yields a unique causal theory

Theorem 6.4 (Reduction to PO)**.** *A CSDP $\alpha=\langle(\mathcal{T},E,\mathsf{X}),D,(U,E)\rangle$ where D is denumerable can be reduced to a PO CSDP.*

Proof. Suppose $D=[M]$ or $D=\mathbb{N}^+$. Take $E'=E\times E^D$ and for $i\in D\cup\{0\}$, $x:=(x_0,x_1,...)\in E'$ define the projection $\mathsf{P}_i(x_0,x_1,...):=x_i$ and the potential outcome variable $\mathsf{X}_i:=\mathsf{X}\circ\mathsf{P}_i$.

Take a map f from $\mathcal{T}$ to causal states on E' such that, letting $(\kappa' F_{\mathsf{X}}, \mu') := f(\kappa, \mu)$, for all $y \in D$ and $A_0, A_1, ... \in \mathcal{E}$:

$$\mu^{po}(A_1 \times ...) := \prod_{y' \in D} \delta_{y'}\kappa(A_{y'}) \tag{10}$$

$$\kappa'(y; A_0 \times A_1 \times ...) := \int_{A_1 \times ...} \delta_{x_y}(A_0)\mu^{po}(dx) \tag{11}$$

$$= \prod_{y' \in D \setminus \{y\}} \delta_{y'}\kappa(A_{y'}) \int_{A_y} \delta_{x_y}(A_0)\delta_y\kappa(dx_y) \tag{12}$$

$$\mu'(A_0 \times A_1 \times ...) := \prod_{y' \in D \setminus \{y\}} \delta_{y'}\kappa(A_{y'}) \int_{A_y} \delta_{x_y}(A_0)\mu(dx_y) \tag{13}$$

It can be verified that κ' is a Markov kernel.

Note that by the definition of conditional probability, for $A, B \in \mathcal{X}$, $\int_{\mathsf{X}_y^{-1}(A)} (\delta_y\kappa')_{|\mathsf{X}_y} F_{\mathsf{X}_0}(x; B)\delta_y\kappa'(dx) = \delta_y\kappa' \curlyvee (F_{\mathsf{X}_0} \otimes F_{\mathsf{X}_y})(A, B)$. Thus by 12, $\delta_x(A)$ is a version of $(\delta_y\kappa')_{|\mathsf{X}_y} F_{\mathsf{X}_0}(x; A)$, so κ' is consequence consistent.

Furthermore, $\mu' F_{\mathsf{X}_y} = \delta_y\kappa F_{\mathsf{X}} = \delta_y\kappa' F_{\mathsf{X}_y}$ for $y \geq 1$. Therefore defining $\mathcal{T}'$ to be the image of $\mathcal{T}$ under f, we can see that $\mathcal{T}'$ is a PO causal theory with respect to "observable" X_0 and "potential outcomes" X_y, $y \in D$.

For $A \in \mathcal{E}$:

$$\kappa' F_{\mathsf{P}_0}(y; A) = \int_E \delta_z(A)\delta_y\kappa(dz) \tag{14}$$

$$= \int_A \kappa(y; dz) \tag{15}$$

$$= \kappa(y; A) \tag{16}$$

For all $B \in \mathcal{X}$

$$\mu' F_{\mathsf{X}_0}(B) = \int_E \delta_z(\mathsf{X}^{-1}(B))\mu(dz) \tag{17}$$

$$= \mu F_{\mathsf{X}}(B) \tag{18}$$

For all $J \in \mathcal{J}$ we have

$$U(\mu F_{\mathsf{X}} J \curlyvee (I_{(D)} \otimes \kappa) = U(\mu' F_{\mathsf{X}_0} J \curlyvee (I_{(D)} \otimes \kappa' F_{\mathsf{P}_0})) \tag{19}$$

$$\tag{20}$$

Therefore, given the PO CSDP $\beta = \langle (\mathcal{T}', E', \mathsf{X}_0), D, (U, E) \rangle$, for all $J \in \mathcal{J}$, $R^\alpha(J, \kappa, \mu) = R^\beta(J, f(\kappa, \mu))$. Thus β is a reduction of α witnessed by f. ☐

Corollary 6.5. *A CBN CSDP for which D is a denumerable set can be reduced to a PO CSDP.*

7 Conclusion

We have shown that CSDPs are an intuitive extension of SDPs and that causal theories that play a fundamental role in CSDPs can naturally represent models posed using the language of CBNs or PO. We believe that causal theories are quite general and capable of representing alternative approaches to causality such as IFMOCS [Peters et al., 2011] or approaches based on group invariance [Besserve et al., 2018].

This perspective raises many questions, for example: 1) Under what conditions do versions of the No-Free Lunch theorems hold for CSDPs? 2) Example 4.3 deals with a crude notion of "continuity" of a causal theory - whether a "nearby" distribution induces a similar risk set, which itelf has implications for learnability of a causal theory. More generally, what properties may be used to characterise the learnability of a causal theory? 3) The notation here borrows heavily from [Fong, 2013], whose diagrammatic representation of Markov kernels is closely related to the DAGs associated with CBNs. Can consequence maps be generically and informatively represented using diagrams similar to DAGs? 4) We have proposed consequence maps and causal theories as "relatively minimal" objects to satisfy the need to connect data, decisions and outcomes. Are there strictly more general objects that may be used instead, and if so under what assumptions are consequence maps and causal theories necessary?

The general perspective proposed in this paper naturally incorporates the two major causal inference frameworks and, for the first time to our knowledge, allows a range of fundamental questions to be formally posed, such as *what are the characteristics of a causal statistical decision problem that make it "learnable"?* Whilst we don't have all the answers, at least we have opened the way to ask such foundational questions!

References

Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms. *arXiv:1901.10912 [cs, stat]*, January 2019. URL `http://arxiv.org/abs/1901.10912`. arXiv: 1901.10912.

Michel Besserve, Bernhard Schoelkopf, Dominik Janzing, et al. Group invariance principles for causal generative models. In *International Conference on Artificial Intelligence and Statistics*, pages 557–565, 2018.

Stephan Bongers, Jonas Peters, Bernhard Schölkopf, and Joris M. Mooij. Theoretical Aspects of Cyclic Structural Causal Models. *arXiv:1611.06221 [cs, stat]*, November 2016. URL `http://arxiv.org/abs/1611.06221`. arXiv: 1611.06221.

Nancy Cartwright. *No Causes in, No Causes out*. Oxford University Press, April 1994. ISBN 978-0-19-159716-9. URL `https://www.oxfordscholarship.com/view/10.1093/0198235070.001.0001/acprof-9780198235071-chapter-3`.

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

A. Philip Dawid. Beware of the DAG! In *Causality: Objectives and Assessment*, pages 59–86, February 2010. URL `http://proceedings.mlr.press/v6/dawid10a.html`.

Philip Dawid. The Decision-Theoretic Approach to Causal Inference. In *Causality*, pages 25–42. John Wiley & Sons, Ltd, 2012. ISBN 978-1-119-94571-0. doi: 10.1002/9781119945710.ch4. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119945710.ch4`.

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.

Sara Geneletti and A Philip Dawid. *Defining and identifying the effect of treatment on the treated*. Citeseer, 2007.

MA Hernán and JM Robins. *Causal Inference*. Chapman & Hall/CRC, 2018.

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`.

Mohammad Ali Mansournia, Julian P. T. Higgins, Jonathan A. C. Sterne, and Miguel A. Hernán. Biases in randomized trials: a conversation between trialists and epidemiologists. *Epidemiology (Cambridge, Mass.)*, 28(1):54–59, January 2017. ISSN 1044-3983. doi: 10.1097/EDE.0000000000000564. URL `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5130591/`.

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

Jonas Peters, Joris M Mooij, Dominik Janzing, and Bernhard Schölkopf. Identifiability of causal graphs using functional models. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, pages 589–598. AUAI Press, 2011.

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

James M Robins and Thomas S Richardson. Alternative graphical causal models and the identification of direct effects. *Causality and psychopathology: Finding the determinants of disorders and their cures*, pages 103–158, 2010.

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`.

Donald B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, 1974. ISSN 1939-2176(Electronic),0022-0663(Print). doi: 10.1037/h0037350.

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 Spirtes, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper, and Thomas Richardson. *Causation, prediction, and search*. MIT press, 2000.

Jin Tian and Judea Pearl. A general identification condition for causal effects. In *Aaai/iaai*, pages 567–573, July 2002.

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.

Supplement to: Causal Statistical Decision Problems

A Markov Kernels

This is an expanded version of Section 2 that explains some notation more thoroughly.

A measurable space $(E, \mathcal{E})$ is a set E and a σ-algebra $\mathcal{E} \subset \mathcal{P}(\mathcal{E})$ containing the measurable sets. A probability measure $\mu \in \Delta(\mathcal{E})$ is a nonnegative map $\mathcal{E} \to [0, 1]$ such that $\mu(\emptyset) = 0$, $\mu(E) = 1$ and for countable $\{E_i\} \in \mathcal{E}$, $\mu(\cup_i E_i) = \sum_i \mu(E_i)$.

We assume all measurable spaces discussed are standard. That is, they are isomorphic to either a subset of $\mathbb{N}$ with the discrete σ-algebra, or $\mathbb{R}$ with the Borel σ-algebra.

Given two measureable sets $(E, \mathcal{E})$ and $(F, \mathcal{F})$, a *Markov kernel* K is a map $E \times \mathcal{F} \to [0, 1]$ where

1. The map $x \mapsto K(x; B)$ is $\mathcal{E}$-measurable for every $B \in \mathcal{F}$
2. The map $B \mapsto K(x; B)$ is a probability measure on $(F, \mathcal{F})$ for every $x \in E$

Abusing notation somewhat, we will give Markov kernels the alternate type signature $K : E \to \Delta(\mathcal{F})$, noting that due to part 1 not every map with this type is a Markov kernel. We will sometimes write the set of Markov kernels of type $E \to \Delta(\mathcal{F})$ as $\Delta(\mathcal{F})^D$, noting again that given part 1, the set of Markov kernels of this type may be smaller than $\Delta(\mathcal{F})^D$.

If we have two random variables $\mathsf{X} : \_ \to X$ and $\mathsf{Y} : \_ \to Y$, the conditional probability $P(\mathsf{Y}|\mathsf{X})$ is a Markov kernel $X \to \Delta(\mathcal{Y})$. Formally, given $\mu \in \Delta(\mathcal{E})$ and a sub-σ-algebra $\mathcal{E}' \subset \mathcal{E}$, there is a Markov kernel $\mu_{|\mathcal{E}'} : E \to \Delta(\mathcal{E})$ such that for $A \in \mathcal{E}$ and $B \in \mathcal{E}'$, $\int_B \mu_{|\mathsf{E}'}(y; A) d\mu(y) = \mu(A \cap B)$. $\mu_{|\mathcal{E}'}$ is a *conditional probability distribution* with respect to $\mathcal{E}'$. This result may not hold if $(E, \mathcal{E})$ is not a standard measureable space [Çinlar, 2011].

Given a set of random variables $\mathbf{X} = \{\mathsf{X}^i\}_{i \in [N]}$ with domain $(E, \mathcal{E})$, $\mu_{|\mathbf{X}} : E \to \Delta(\mathcal{E})$ is a conditional probability distribution with respect to the σ-algebra generated by $\mathbf{X}$: $\sigma(\cup_{i \in [N]} \sigma(\mathcal{X}^i))$. We will use this subscript notation rather than the more common bar notation (e.g. $\mu(\cdot|\mathbf{X})$) to express conditional probability from here onwards.

Two Markov kernels $K : E \to \Delta(\mathcal{F})$ and $K' : E \to \Delta(\mathcal{F})$ are μ-almost surely equivalent given $\mu \in \Delta(\mathcal{E})$ if

$$\int_A K(x; B) d\mu = \int_A K'(x; B) d\mu \qquad \forall A \in \mathcal{E}, B \in \mathcal{F} \tag{21}$$

A.1 Operations with Markov kernels

For the following, assume K is a Markov kernel from $E \to \Delta(\mathcal{F})$, K' a kernel $E \to \Delta(\mathcal{H})$, L is a Markov kernel $F \to \Delta(\mathcal{G})$, μ is a probability measure on $(E, \mathcal{E})$, ν is a probability measure on $(F, \mathcal{F})$ and f is a nonnegative measurable function $F \to \mathbb{F}$.

The notation here borrows heavily from Çinlar [2011] and Fong [2013].

A.1.1 Kernel products

The kernel-kernel product KL is a Markov kernel $E \to \Delta(\mathcal{G})$ such that $KL(x; B) := \int_F K(x; dy) L(y; B), \qquad x \in E, B \in \mathcal{G}$.

The measure-kernel product of μ and K, μK is a probability measure on $(F, \mathcal{F})$ such that $\mu K(B) = \int_E \mu(dx) K(x; B), \qquad B \in \mathcal{F}$.

The kernel-function product Kf is a nonnegative measurable function $E \to \mathbb{R}$ such that $Kf(x) := \int_F K(x; dy) f(y), \qquad x \in E$.

Kernel products are in general associative: $(KL)M = K(LM)$.

A.1.2 Special kernels

$I_{(E)}$ is a kernel $E \to \Delta(\mathcal{E})$ defined by $x \mapsto \delta_x$. It has the properties $\mu I_{(E)} = \mu$, $KI_{(F)} = K$, $I_{(E)} K = K$, $I_{(F)} f = f$.

$\curlyvee_E$ is a kernel $E \to \Delta(\mathcal{E} \otimes \mathcal{E})$ defined by $x \mapsto \delta_{(x,x)}$. We will subsequently leave the space implicit. The symbol $\curlyvee$ is pronounced "splitter".

Given $M : H \to \Delta(\mathcal{I})$, $K \otimes M$ is a Markov kernel $E \times H \to \Delta(\mathcal{F} \otimes \mathcal{I})$ where

$$K \otimes M(x, y; A \times B) := K(x; A)M(y; B) \tag{22}$$

Given $N : I \to \Delta(\mathcal{J})$, it can be verified that $(K \otimes M)(L \otimes N) = KL \otimes MN$.

$\curlyvee(K \otimes K')$ is a Markov kernel $E \to \Delta(\mathcal{F} \otimes \mathcal{H})$ and

$$\begin{aligned}
\curlyvee(K \otimes K')(x; A \times B) &= \int_E K(x'; A)K'(x''; B)\delta_{(x,x)}(dx' \times dx'') &(23)\\
&= K(x; A)K'(x; B) &(24)
\end{aligned}$$

We can overload notation to use $\curlyvee(K \otimes K' \otimes K'')$ for the nested construction $\curlyvee(K \otimes \curlyvee(K' \otimes K''))$.

Let $(*, \{\emptyset, *\})$ be an indiscrete measurable set. $\upharpoonleft_E$ is a kernel $E \to \Delta(\{\emptyset, *\})$ defined by $x \mapsto \mathbb{1}_*$. We have $\curlyvee(I \otimes \upharpoonleft) = I$. The symbol $\upharpoonleft$ is pronounced "stopper".

Given some measurable function $g : E \to F$, the kernel $F_g : E \to \Delta(\mathcal{F})$ is defined by $x \mapsto \delta_{g(x)}$. It is easy to check that $F_g F_g = F_g$. For $\mu \in \Delta(\mathcal{E})$, the product μF_g is the push forward measure $g_*\mu$.

$$\begin{aligned}
\mu F_g(A) &= \int_E \delta_{g(x)}(A)d\mu &(25)\\
&= \mu(g^{-1}(A)) &(26)\\
&= g_*\mu(A) &(27)
\end{aligned}$$

Given two random variables $\mathsf{X} : (E, \mathcal{E}) \to (X, \mathcal{X})$ and $\mathsf{Y} : (E, \mathcal{E}) \to (Y, \mathcal{Y})$, the product $\mu\curlyvee(F_\mathsf{X} \otimes F_\mathsf{Y})$ is the joint distribution of X and Y.

$$\begin{aligned}
\mu\curlyvee(F_\mathsf{X} \otimes F_\mathsf{Y})(A, B) &= \int_E \delta_{\mathsf{X}(x)}(A)\delta_{\mathsf{Y}(x)}(B)d\mu &(28)\\
&= \mu(\mathsf{X}^{-1}(A) \cap \mathsf{Y}^{-1}(B)) &(29)
\end{aligned}$$

B Appendix: Causal Statistical Decision Problems

Lemma B.1 (Reduction preserves admissibility)**.** *If a CSDP β with induced game $\langle \mathcal{T}, \mathcal{J}, R\rangle$ can be reduced to a statistical decision problem α with induced game $\langle \mathcal{H}, \mathcal{J}, R'\rangle$ then a decision function $J \in \mathcal{J}$ is admissible in β iff it is admissible in α.*

Proof. Suppose $J \in \mathcal{J}$ is inadmissible in α. Then there is some $J' \in \mathcal{J}$, $\mu \in \mathcal{H}$ such that $R'(J', \mu) < R'(J, \mu)$ and $R'(J', \nu) \leq R'(J, \nu)$ for all $\nu \in \mathcal{H}$. Let h be the function that witnesses the reduction. Then we have for all $\tau \in h^{-1}(\mu)$, $R(J', \tau) = R'(J', \mu) < R(J, \tau) = R'(J, \nu)$ and for all $\nu \in \mathcal{H}$, $\chi \in h^{-1}(\nu)$, $R(J', \chi) = R'(J', \nu) \leq R(J, \chi) = R'(J, \nu)$. The set $\bigcup_{\nu\in\mathcal{H}} h^{-1}(\nu) = \mathcal{T}$, so J is inadmissible in β.

Suppose $J \in \mathcal{J}$ is admissible in β. Then there is some $J' \in \mathcal{J}$, $\tau \in \mathcal{T}$ such that $R(J', \tau) < R(J, \tau)$ and $R(J', \chi) \leq R(J, \chi)$ for all $\chi \in \mathcal{T}$. Then we have $R'(J', h(\tau)) = R(J', \tau) < R(J, \tau) = R'(J, h(\tau))$ and $R'(J', h(\chi)) = R(J', \chi) \leq R(J, \chi) = R'(J, h(\chi))$. Because h is surjective, J is admissible in α. ☐

Corollary B.2 (Reduction preserves completeness)**.** *If a causal decision problem β with induced game $\langle \mathcal{T}, \mathcal{J}, R\rangle$ can be reduced to a statistical decision problem α with induced game $\langle \mathcal{H}, \mathcal{J}, R'\rangle$, then an (essentially) complete class with respect to α is (essentially) complete with respect to β.*

Lemma B.3 (Induced Bayes rule)**.** *If a CSDP β with induced game $\langle \mathcal{T}, \mathcal{J}, R\rangle$ can be reduced to a statistical decision problem α with induced game $\langle \mathcal{H}, \mathcal{J}, R'\rangle$ witnessed by $h : \mathcal{T} \to \mathcal{H}$ and $J^{\xi}_{ba} \in \mathcal{J}$ is a Bayes rule with respect to the problem α and the prior ξ then J^{ξ}_{ba} is a Bayes rule with respect to the problem β and the induced prior ξ_h.*

Proof. For any $J \in \mathcal{J}$, $\tau \in \mathcal{T}$, by the properties of the push-forward measure

$$\int_{\mathcal{T}} R(J, \tau) d\xi_h = \int_{\mathcal{H}} R'(J, h(\tau)) d\xi \tag{30}$$

And therefore, if a Bayes rule exists,

$$\underset{J\in\mathcal{J}}{\arg\min} \int_{\mathcal{T}} R(J, \tau) d\xi_h = \underset{J\in\mathcal{J}}{\arg\min} \int_{\mathcal{H}} R'(J, h(\tau) d\xi \tag{31}$$

☐

Theorem B.4 (Complete class theorem (CSDP))**.** *Given an CSDP $\alpha := \langle(\mathcal{T}, E), D, \mathsf{X}, U\rangle$ with risk R, if there is a reduction to an SDP $\beta := \langle(\mathcal{H}, F), D, \mathsf{Y}, \ell\rangle$ with risk R' such that $|\mathcal{H}| < \infty$ and $\inf_{J\in\mathcal{J},\mu\in\mathcal{H}} R'(J, \mu) < -\infty$ then the set of all Bayes decision functions is a complete class and the set of all admissible Bayes decision functions is a minimal complete class.*

Proof. Given the conditions, the Bayes decision functions in β form a complete class and admissible Bayes rules a minimal complete class [Ferguson, 1967].

By Corollary B.2 the Bayes rules for β are complete in α, and the admissible Bayes rules for β are essentially complete in α.

Every (admissible) Bayes rule for β is a(n admissible) Bayes rule for α, so the set of (admissible) Bayes rules for α is also (essentially) complete in α. ☐

Theorem B.5 (Reduction of a CSDP on observations)**.** *A CSDP $\alpha = \langle(\mathcal{T}^{\alpha}, (E, \mathcal{E}), \mathsf{X}), D, (U, (F, \mathcal{F}))\rangle$ where, for $\zeta \in \Delta(\mathcal{E} \otimes \mathcal{D})$ can be reduced to a problem $\beta = \langle(\mathcal{T}^{\beta}, (X, \mathcal{X}), \mathrm{id}_X), D, (U, (F, \mathcal{F})\rangle$ by marginalization.*

Proof. Consider the mapping $g : \mathcal{T}^{\alpha} \to \mathcal{T}^{\beta}$ given by $(\kappa, \mu) \mapsto (\kappa, \mu F_{\mathsf{X}})$.

For $J \in \mathcal{J}$, $(\kappa, \mu) \in \mathcal{T}^\alpha$

$$
\begin{aligned}
R^\alpha(J,\kappa,\mu) &= \sup_{\gamma' \in \Delta(\mathcal{D})} U(\gamma'^{\curlyvee}(I_{(D)} \otimes \kappa)) - U(\mu F_{\mathsf{X}} J^{\curlyvee}(I_{(D)} \otimes \kappa)) && (32)\\
&= \sup_{\gamma' \in \Delta(\mathcal{D})} U(\gamma'^{\curlyvee}(I_{(D)} \otimes \kappa) - U(\mu F_{\mathsf{X}} F_{\mathsf{X}} J^{\curlyvee}(I_{(D)} \otimes \kappa))) && (33)\\
&= R^\beta(J, g(\kappa,\mu)) && (34)
\end{aligned}
$$

□

Theorem B.6 (Reduction of a CSDP on the utilty)**.** *Given a CSDP* $\alpha = \langle(\mathcal{T}^\alpha, (E,\mathcal{E}), \mathsf{X}), D, (U,(F,\mathcal{F}))\rangle$ *where, for* $\zeta \in \Delta(\mathcal{E} \otimes \mathcal{D})$*, if* $U(\zeta) = U'(\zeta(I_{(D)} \otimes F_{\mathsf{Y}})))$ *for some* $\mathsf{Y} : F \to Y$ *and* $U' : \Delta(\mathcal{Y}) \to \mathbb{R}$ *then* α *has* Y-observable utility. *Such a problem can be reduced to a problem* $\beta = \langle(\mathcal{T}^\beta, (E,\mathcal{E}), \mathsf{X}), D, (U', (Y,\mathcal{Y}))\rangle$ *by marginalization.*

Proof. Consider the mapping $g : \mathcal{T}^\alpha \to \mathcal{T}^\beta$ given by $(\kappa,\mu) \mapsto (\kappa F_{\mathsf{Y}}, \mu)$.

We have for $J \in \mathcal{J}$, $(\kappa, \mu) \in \mathcal{T}^\alpha$

$$
\begin{aligned}
R^\alpha(J,\kappa,\mu) &= \sup_{\gamma' \in \Delta(\mathcal{D})} U(\gamma'^{\curlyvee}(I_{(D)} \otimes \kappa)) - U(\mu F_{\mathsf{X}} J^{\curlyvee}(I_{(D)} \otimes \kappa)) && (35)\\
&= \sup_{\gamma' \in \Delta(\mathcal{D})} U'(\gamma'^{\curlyvee}(I_{(D)} \otimes \kappa)(I_{(D)} \otimes F_{\mathsf{Y}})) - U'(\mu F_{\mathsf{X}} J^{\curlyvee}(I_{(D)} \otimes \kappa))(I_{\mathsf{D}} \otimes F_{\mathsf{Y}})) && (36)\\
&= \sup_{\gamma' \in \Delta(\mathcal{D})} U'(\gamma'^{\curlyvee}(I_{(D)} \otimes \kappa F_{\mathsf{Y}})) - U'(\mu F_{\mathsf{X}} J^{\curlyvee}(I_{(D)} \otimes \kappa F_{\mathsf{Y}})) && (37)\\
&= R^\beta(J, g(\kappa,\mu)) && (38)
\end{aligned}
$$

□

Theorem B.7. *Every SDP* $\langle(\mathcal{H}, E, \mathsf{X}), D, \ell\rangle$ *can be reduced to a CSDP.*

Proof. Take D to be the projection from $D \times E$ to D. For each $\mu \in \mathcal{H}$ define the consequence $\kappa_\mu : d \mapsto \mu$ for all $d \in D$. Take the causal theory $\mathcal{T} = \{(\kappa_\mu, \mu) | \mu \in \mathcal{H}\}$ for some $\pi \in \Delta(\mathcal{D})$ and the pseudo-utility $U(\nu) = -\mathbb{E}_\nu\left[\ell(P^\nu_{\mathsf{E}}, \mathsf{D})\right]$ to construct the CSDP $\langle(\mathcal{T}, E, \mathsf{X}), D, (U, E)\rangle$. We will show that the original problem can be reduced to this.

For $\gamma \in \Delta(\mathcal{D})$ the induced loss L is

$$
\begin{aligned}
L(\kappa_\mu,\gamma) &= -\sup_{\gamma' \in \Delta(\mathcal{D})} \mathbb{E}_{\gamma'^{\curlyvee}(I_{(D)} \otimes \kappa_\mu)_{|\mathsf{E}}}[\ell(\gamma'^{\curlyvee}(I_{(D)} \otimes \kappa_\mu)_{|\mathsf{E}}, \mathsf{D})] + \mathbb{E}_{\gamma^{\curlyvee}(I_{(D)} \otimes \kappa_\mu)}[\ell(\gamma^{\curlyvee}(I_{(D)} \otimes \kappa_\mu)_{|\mathsf{E}}, \mathsf{D})] && (39)\\
&= \mathbb{E}_\gamma[\ell(\mu, \mathsf{D})] && (40)
\end{aligned}
$$

For the surjective map, take $g : \mathcal{H} \to \mathcal{T}$ defined by $g(\mu) = \kappa_\mu$.

Denote by R the risk associated with the SDP $\langle(\mathcal{H}, E), D, \mathsf{X}, \ell\rangle$ and by R' the risk associated with the CSDP $\langle(\mathcal{T}, E), D, \mathsf{X}, U\rangle$. Then

$$
\begin{aligned}
R'(J,\kappa,\mu) &= \int_D \ell(\mu, y)\mu F_{\mathsf{X}} J(dy) && (41)\\
&= R(J, g(\kappa,\mu)) && (42)
\end{aligned}
$$

□

Theorem B.8. *Given a CSDP* $\beta = \langle(\mathcal{T}, E, \mathsf{X}), D, (U, F)\rangle$ *where* U *is an ordinary pseudo-utility, let* $\mathcal{K} = \{\kappa | (\kappa,\mu) \in \mathcal{T}\}$ *be the set of consequences.* β *is reducible to a statistical decision problem on the measurable space* $(E \times F \times D, \mathcal{E} \otimes \mathcal{F} \otimes \mathcal{D})$ *if there is some surjective map* $m : \Delta(\mathcal{F} \otimes \mathcal{D}) \to \mathcal{K}$.

Proof. Let $\mathcal{H} \subset \Delta(\mathcal{E} \otimes \mathcal{F} \otimes \mathcal{D})$ be some hypothesis class and let $m^\dagger$ be a right inverse of m. Define $h : \mathcal{T} \to \mathcal{H}$ by $(\kappa, \mu) \mapsto \mu \otimes m^\dagger(\kappa)$.

Let $k : \Delta(\mathcal{F})^D \times D \to \mathbb{R}$ be the differential loss induced by the ordinary pseudo-utility U (see Equation 3).

Given the projections $\mathsf{F} : E \times F \times D \to F$ and $\mathsf{D} : E \times F \times D \to D$ and arbitrary $\xi \in \Delta(\mathcal{E} \otimes \mathcal{F} \otimes \mathcal{D})$ define $\ell : \mathcal{H} \times D \to [0, \infty)$ by

$$\ell(\xi, y) = k(m(\xi F_{\curlyvee (\mathsf{F} \otimes \mathsf{D})}), y) \tag{43}$$

Note that

$$\ell(h(\kappa, \mu), y) = k(\kappa, y) \tag{44}$$

Define $\mathsf{X}' : E \times F \times D \to X$ by $(a, b, c) \mapsto \mathsf{X}(a)$.

Then, given the statistical decision problem $\langle(\mathcal{H}, E \times F \times D, \mathsf{X}'), D, \ell\rangle$, we have for all $J \in \mathcal{J}$, $(\kappa, \mu) \in \mathcal{T}$ the risk

$$\begin{aligned}
R'(J, h(\kappa, \mu)) &= \int_D \ell(h(\kappa, \mu), y) h(\kappa, \mu) F_{\mathsf{X}'} J(dy) &(45)\\
&= \int_D \ell(h(\kappa, \mu), y)(\mu \otimes m^\dagger(\kappa)) F_{\mathsf{X}'} J(dy) &(46)\\
&= \int_D k(\kappa, y) \mu F_{\mathsf{X}} J(dy) &(47)\\
&= R(J, \kappa, \mu) &(48)
\end{aligned}$$

$\square$

Example B.9 (Irreducible CSDP)**.** The choice of decision function in an SDP does not affect the state, while this choice does affect the outcome in an CSDP. For an SDP, then, the risk of a mixed decision function is equal to the mixture of risks of each atomic decision function but this is not true in general for an CSDP.

Take the CSDP $\langle(\mathcal{T}, E), D, \mathsf{X}, U\rangle$ where $E = D = \{0, 1\}$, $\mathsf{Y} : E \to \{0, 1\}$ is the identity function, $U : \mu \mapsto -\mathrm{Var}_\mu[\mathsf{Y}]$ and $\mathcal{T} = \{(d \mapsto \delta_d, \nu) | \nu \in \Delta(\mathcal{E})\}$.

For any $(\kappa, \mu) \in \mathcal{T}$ and $J \in \mathcal{J}$ we have

$$R(J, \kappa, \mu) = 0.25 - \mathrm{Var}_{\mu F_{\mathsf{X}} J}(\mathsf{Y}) \tag{49}$$

Consider the forgetful decision functions $J_0 : x \mapsto$ Bernoulli(0) and $J_{1/2} : x \mapsto \mathrm{Bernoulli}(\frac{1}{2})$ and $J_1 : x \mapsto \mathrm{Bernoulli}(1)$ for all $x \in X$. Note that $J_{1/2}(x; A) = \frac{1}{2}(J_0(x; A) + J_1(x; A))$ for all $x \in X$, $A \in \mathcal{D}$. For any statistical decision problem with risk R',

$$\begin{aligned}
R'(J_{1/2}, \mu) &= \int_D \ell(\mu, y) \mu F_{\mathsf{X}} J_{1/2}(dy) &(50)\\
&= \frac{1}{2}\left(\int_D \ell(\mu, y) \mu F_{\mathsf{X}} J_0(dy) + \int_D \ell(\mu, y) \mu F_{\mathsf{X}} J_1(dy)\right) \quad = \frac{1}{2}\left(R'(J_0, \mu) + R'(J_1, \mu)\right) &(51)
\end{aligned}$$

But

$$\begin{aligned}
R(J_{1/2}, \kappa, \mu) &= 0 &(52)\\
&\neq \frac{1}{2}\left(R(J_0, \kappa, \mu) + R(J_1, \kappa, \mu)\right) &(53)
\end{aligned}$$

Corollary B.10. *The class of nonrandomized decision functions is not essentially complete for CSDPs. The stochastic decision function $J_{1/2}$ is strictly better than any deterministic function in the above example.*

C Appendix: CBN is a causal theory

Theorem C.1. *Given a measurable set $(E, \mathcal{E})$ and a graph $\mathcal{G}$ over a set of random variables $\{\mathsf{X}^i\}_{i\in[N]}$ where $\mathsf{X}^i : E \to X^i$, a decision set $(D, \mathcal{D})$ and random variables $\{\mathsf{D}^i\}_{i\in[N]}$ with $\mathsf{D}^i : (D, \mathcal{D}) \to (X^i \cup \{*\}, \sigma(\mathcal{X}^i \cup \{*\})$. Given $\mu \in \mathcal{G}_{\mathcal{G}}$, let μ^y be the $\mathcal{G}, \mu, y$-interventional distribution (Definition 4.2).*

Then the map $\kappa^{\mu,\mathcal{G}} : D \to \Delta(\mathcal{E})$ given by $y \mapsto \mu^y$ is a Markov kernel with respect to $(D, \mathcal{D})$ and $(E, \mathcal{E})$.

Proof. The DAG $\mathcal{G}$ induces a partial ordering on the RV's X^i by $\mathsf{X}^i < \mathsf{X}^j$ if $\mathsf{X}^i \to \mathsf{X}^j$ is in $\mathcal{G}$. Without loss of generality, suppose the total ordering $X^0, ..., X^N$ is consistent with the partial ordering induced by $\mathcal{G}$.

Let $\kappa^i : \mathcal{E} \to \Delta(\mathcal{X}^i)$ be defined by $\kappa^i(x; A) := \mu_{|\mathsf{X}^{<i}} F_{\mathsf{X}^i}$. Note that by the compatibility of μ, for all $x \in \mathcal{E}$, $A \in \mathcal{X}^i$ we also have

$$\kappa^i(x; A) = \mu_{|\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)} F_{\mathsf{X}^i}(x; A) \tag{54}$$

Consider $\kappa^{i,*} : D \times E \to \Delta(\mathcal{X}^i)$ given by

$$\kappa^{i,*}(y, pa^i; A) := \begin{cases} \kappa^i(pa^i; A) & \mathsf{D}^i(y) = * \\ \delta_{\mathsf{D}^i(y)}(A) & \mathsf{D}^i(y) \neq * \end{cases} \tag{55}$$

Clearly for every $(d, pa^i) \in D \times E$ the map $A \mapsto \kappa^{i,*}(d, pa^i; A)$ is a probability distribution on $\mathcal{X}^i$. Fix $B \in \mathcal{X}_i$ and let $\kappa_B^{i,*} = \kappa_i'(\cdot; B)$.

Then for any $A \in \mathcal{B}([0,1])$

$$\begin{aligned}
[\kappa_B^{i,*}]^{-1}(A) &= [\mathsf{D}^i]^{-1}(\{*\}) \times [\kappa_i^B]^{-1}(A) & \text{if } 0, 1 \notin A \quad (56)\\
&= [\mathsf{D}^i]^{-1}(\{*\}) \times [\kappa_i^B]^{-1}(A) \cup [\mathsf{D}^i]^{-1}(B) \times X^{\mathrm{Pa}_{\mathcal{G}}(i)} & \text{if } 1 \in A \wedge 0 \notin A \quad (57)\\
&= [\mathsf{D}^i]^{-1}(\{*\}) \times [\kappa_i^B]^{-1}(A) \cup [\mathsf{D}^i]^{-1}(B^C) \times X^{\mathrm{Pa}_{\mathcal{G}}(i)} & \text{if } 0 \in A \wedge 1 \notin A \quad (58)\\
&= [\mathsf{D}^i]^{-1}(\{*\}) \times [\kappa_i^B]^{-1}(A) \cup [\mathsf{D}^i]^{-1}(X^i) \times X^{\mathrm{Pa}_{\mathcal{G}}(i)} & \text{if } 0 \in A \wedge 1 \in A \quad (59)
\end{aligned}$$

Note that $\sigma(\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i) \subset \mathcal{E}$ and $[\kappa_i^B]^{-1}(A) \in \sigma(\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)$. Further note that $\{*\}$, B and B^C are in $\sigma(\mathcal{X}^i \cup \{*\})$. Therefore, in every case the result is an element of $\mathcal{E} \otimes \mathcal{D}$ and $\kappa^{i,*}$ is a Markov kernel.

Then $\iota^{\mathcal{G}} : D \to \Delta(\mathcal{X})$ defined below is a Markov kernel.

$$\iota^{\mathcal{G}} : (y; A) \mapsto \int_{A^0} \kappa^{0,*}(y; dx^0) ... \int_{A^{N-1}} \kappa^{N-1,*}(y, x^{n-2}; dx^{n-1}) \kappa^{N,*}(y, x^{n-1}; A^N) \tag{60}$$

for $y \in D$, $A \in E$ and $A^i = [\mathsf{X}^i]^{-1}(A)$.

From Equations 54, 55 and 60 we can verify that, given some $i \in N$, if $\mathsf{D}^i(y) = \{*\}$ then $[\delta_y \iota^{\mathcal{G}}]_{\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)} = \kappa_i = \mu_{\mathrm{Pa}_{\mathcal{G}}(\mathsf{X}^i)} F_{\mathsf{X}^i}$ and if $\mathsf{D}^i(y) \neq \{*\}$ then $\delta_y \iota^{\mathcal{G}} = \delta_{\mathsf{D}^i(y)} F_{\mathsf{X}^i}$. From Equation 60 and the compatibility of μ with $\mathcal{G}$ it further follows that $\delta_y \iota^{\mathcal{G}}$ is compatible with $\mathcal{G}$. Therefore $\delta_y \iota^{\mathcal{G}} = \mu^y$ and so $\iota^{\mathcal{G}} = \kappa^{\mu,\mathcal{G}}$. $\square$

D Appendix: Counterfactuals

A causal theory for Potential Outcomes is associated with a much larger hypothesis class than any causal theory that works only with distributions over observable variables. Theorems B.5 and B.6 show that given any SCDP based on Potential Outcomes, provided that the potential outcome variables are unobserved and the utility does not depend on them, a reduced SCDP can be constructed by marginalising over potential outcomes. Potential outcomes are not universally excluded by this; there are some examples of problems where one does care about the values of potential outcome variables. The *effect of treatment on the treated* (ETT) that depends on counterfactual quantities and has some relevance to decision preferences Rubin [1974], though it is controversial whether this dependence is necessary Geneletti and Dawid [2007]. More straightforwardly, the legal standard of "no harm but for the defendant's negligence" does seem to invoke fundamentally counterfactual considerations Pearl [2009].

Example D.1 (Performance bias)**.** Suppose we have a CSDP $\langle(\mathcal{T}, E), D, \mathsf{X}, (U, E)\rangle$ where the observed data X is from a randomised controlled trial (RCT), $\mathsf{Y}_0 : E \to Y$ and $\mathsf{Y}_1 : E \to Y$ are random variables representing a particular outcome of interest under no treatment and treatment respectively and $\mathsf{Y} : E \to Y$ represents the "realised" outcome of interest and for $\xi \in \Delta(\mathcal{E})$, $U(\xi) = \mathbb{E}_\xi[\mathsf{Y}]$.

Under usual assumptions about RCTs, if we suppose the observed data are distributed according to $\mu \in \Delta(\mathcal{E})$ it is possible (given infinite data X) to determine $\mathbb{E}_\mu[\mathsf{Y}_0]$ and $\mathbb{E}_\mu[\mathsf{Y}_1]$ [Rubin, 2005].

Consequence consistency is assumed, but performance bias is suspected, which can lead to $\delta_i\kappa\mathsf{Y}_i$ differing from $\mathbb{E}_\mu[\mathsf{Y}_i]$ [Mansournia et al., 2017].

1. Assume performance bias is absent, so the theory must satisfy $\delta_i\kappa\mathsf{Y}_i = \mathbb{E}_\mu[\mathsf{Y}_i]$
2. Assume performance bias has a uniform additive effect: the theory satisfies $\delta_i\kappa\mathsf{Y}_i = \mathbb{E}_\mu[\mathsf{Y}_i] + k$. In this case the average treatment effect can still be estimated from the data: $\delta_1\kappa\mathsf{Y}_1 - \delta_0\kappa\mathsf{Y}_0 = \mathbb{E}[\mathsf{Y}_1] - \mathbb{E}[\mathsf{Y}_0]$ which may be sufficient to find a decision function minimising the risk
3. Avoid assumptions about the effect of performance bias; the theory satisfies no particular relationship between $\mathbb{E}_\mu[\mathsf{Y}_i]$ and $\delta_i\kappa\mathsf{Y}_i$ and we may therefore expect preferred decision function to ignore the data

The question of specifying this relationship arises naturally when we consider connecting Potential Outcomes to CSDPs. Nonetheless, the possibility of deviations from option 1 above are often treated as "external to the causal problem". For example, Mansournia et al. [2017] states:

> In this case, it might be more appropriate to say that the intention-to-treat effect from the trial is not generalizable or transportable to other settings rather than saying that it is "biased"