Understanding Causal Primitives Using Modular Probability

David Johnston

September 22, 2021

1 Introduction

Two widely used approaches to causal modelling are *graphical causal models* and *potential outcomes models*. Graphical causal models, which include Causal Bayesian Networks and Structural Causal Models, provide a set of *intervention* operations that take probability distributions and a graph and return a modified probability distribution (Pearl, 2009). Potential outcomes models feature *potential outcome variables* that represent the "potential" value that a quantity of interest would take under the right circumstances, a potential that may be realised if the circumstances actually arise, but will otherwise remain only a potential or *counterfactual* value (Rubin, 2005).

One challenge for both of these approaches is understanding how their causal primitives – interventions and potential outcome variables respectively – relate to the causal questions we are interested in. This challenge is related to the distinction, first drawn by (Korzybski, 1933), between "the map" and "the territory". Causal models, like other models, are "maps" that purport to represent a "territory" that we are interested in understanding. Causal primitives are elements of the maps, and the things to which they refer are parts of the territory. The maps contain all the things that we can talk about unambiguously, so it is challenging to speak clearly about how parts of the maps relate to parts of the territory that fall outside of the maps.

For example, Hernán and Taubman (2008), who observed that many epidemiological papers have been published estimating the "causal effect" of body mass index and argued that, because *actions* affecting body mass index[1] are vaguely defined, potential outcome variables and causal effects themselves become ill-defined. We note that "actions targeting body mass index" are not elements of a potential outcomes model but "things to which potential outcomes should correspond". The authors claim is that vagueness in the "territory" leads to ambiguity about elements of the "map" – and, as we have suggested, anything we can try to say about the territory is unavoidably vague. This seems like a serious problem.

[1] the authors use the term "intervention", but they do not use it mean a formal operation on a graphical causal model, and we reserve the term for such operations to reduce ambiguity.

In a response, Pearl (2018) argues that *interventions* (by which we mean the operation defined by a causal graphical model) are well defined, but may not always be a good model of an action. Pearl further suggests that interventions in graphical models correspond to "virtual interventions" or "ideal, atomic interventions", and that perhaps carefully chosen interventions can be good models of actions. Shahar (2009), also in response, argued that interventions targeting body mass index applied to correctly specified graphical causal models will necessarily yield no effect on anything else which, together with Pearl's suggestion, implies perhaps that an "ideal, atomic intervention" on body mass index cannot have any effect on anything else. If this is so, it seems that we are dealing with quite a serious case of vagueness – there is a whole body of literature devoted to estimating a "causal effect" that, it is claimed, is necessarily equal to zero! Authors of the original literature on the effects of BMI might counter that they were estimating something different that wasn't necessarily zero, but as far as we are concerned such a response would only underscore the problem of ambiguity.

One of the key problems in this whole discussion is how the things we have called *interventions* – which are elements of causal models – relate to the things we have called *actions*, which live outside of causal models. One way to address this difficulty is to construct a bigger causal model that can contain both "interventions" and "actions", and we can then speak unambiguously about how one relates to another. This is precisely what we do here.

To do this, we use a novel approach to probability modelling that we find is well suited to building causal mdoels. A typical approach to probability modelling is to construct a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ that serves as a top level model, along with a collection of random variables defined by measurable functions on this space, such that the particular quantities of interest can be obtained from conditional and marginal distributions on this space. Instead we consider a modelling context $\mathcal{M}$ that contains a collection of *probability components*, which are Markov kernels with named inputs and outputs. The names correspond to variables in the standard setting. Probability components with the right input and output types can be *connected*, an operation that yields a new probability component. We relate this back to the standard approach by equipping each probability component with a probability space and requiring that all components are the conditional probability distributions on their assigned spaces corresponding to their input and output labels.

Equipped with this foundation, we apply it to a variety of approaches to causal modelling, showing how it can enable understanding of different approaches in a common framework, and how it can represent assertions that were previously made "outside the model". First, we consider causal decision problems and derive *see-do models*, which reduce to statistical decision problems when augmented with the principle of expected utility. See-do models are a particular kind of probability component that we call a *comb*, which can be thought of as a probability model that needs something to be inserted into the middle. We consider causal graphical models, and show how under a very slight modification to the standard notation they induce see-do models, which allows us to formally connect *interventions* to *actions*. Finally, we consider potential out-

comes models and show how we can formalise the typical assertion (which again, lives "outside the model") that potential outcomes represent counterfactual values. Potential outcomes models as typically used do not contain counterfactual assertions and in fact feature comb and insert components almost but not quite identical to combs and inserts found in causal graphical models.

I'm probably going to have to cut some of the above

Contents

2 Technical prerequesites

Our theory makes heavy use of *Markov kernels* or *stochastic functions*, which are taken from probability theory. However, the manner in which we use them is non-standard. The usual way to apply probability theory to model building is to assume we have a probability space $(\mathbb{P}, \Omega, \mathcal{F})$ with random variables defined as functions with domain Ω, and all aspects of the model of interest are supposed

to be captured by this. Under our approach, we instead consider components, represented by Markov kernels $\mathbf{K} : E \to \Delta(F)$ along with labeled inputs and outputs. The labels do the same job that random variables do in the usual formulation. These components can be composed or broken apart, but we do not assume that there is an overarching probability space from which all components can be derived.

In addition, we introduce a graphical notation for Markov kernels that is the subject of a coherence theorem: two Markov kernels represented by pictures that differ only by planar deformations are identical (Selinger, 2010).

2.1 Markov kernels

Markov kernels can be thought of as measurable functions that map to probability distributions. A conditional probability $\mathbb{P}(\mathsf{Y}|\mathsf{X})$, which maps from values of X to probability distributions over Y, and an interventional map $x \mapsto \mathbb{P}(\mathsf{Y}|do(\mathsf{X} = x))$ that likewise maps values of X to probability distributions on Y, are both Markov kernels.

Our theory is susbtantially simplified by restricting our attention to discrete sets – that is, sets X with at most a countable number of elements endowed with the σ-algebra made up of every subset of X, also called the discrete σ-algebra.

In the discrete setting, we can represent probability distributions as covectors, Markov kernels as matrices and measurable functions as vectors.

Given a set X, a probability distribution $\mathbb{P}$ on X is a covector in $\mathbb{R}^{|X|}$, which we will write $\mathbb{P} := (\mathbb{P}^i)_{i \in X}$. To be a probability distribution we require

$$0 \leq P_i \leq 1 \qquad \forall i \in X \tag{1}$$

$$\sum_i P_i = 1 \tag{2}$$

Given discrete sets X and Y, a Markov kernel $\mathbf{K} : X \to \Delta(Y)$ is a matrix in $\mathbb{R}^{|X| \times |Y|}$; $\mathbf{K} = (K_i^j)_{i \in X, j \in Y}$ where

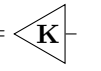
$$0 \leq K_i^j \leq 1 \qquad \forall i, j \tag{3}$$

$$\sum_{i \in X} K_i^j = 1 \qquad \forall j \tag{4}$$

Rows of Markov kernel are probability distributions: $\mathbf{K}_x := (K_x^j)_{j \in Y}$. Alternatively, we can consider probability distributions to be Markov kernels with one row.

Graphically, we represent a Markov kernel as a box and a probability distribution as a triangle:

$\mathbf{K} :=$  (5)

$\mathbb{P} :=$  (6)

2.2 Cartesian and tensor products

The Cartesian product $X \times Y := \{(x, y) | x \in X, y \in Y\}$.

Given kernels $\mathbf{K} : W \to Y$ and $\mathbf{L} : X \to Z$, the tensor product $\mathbf{K} \otimes \mathbf{L} : W \times X \to \Delta(Y \times Z)$ is defined by $(\mathbf{K} \otimes \mathbf{L})^{(y,z)}_{(w,x)} := K^y_w L^z_x$.

Graphically, the tensor product is represeted by parallel juxtaposition:

$$\mathbf{K} \otimes \mathbf{L} := \tag{7}$$

2.3 Delta measures, erase maps, copy maps

The iverson bracket $[\![\cdot]\!]$ evaluates to 1 if $\cdot$ is true and 0 otherwise.

For any X and any $x \in X$, $\delta[x]$ is the probability measure defined by $\delta[x]^i = [\![x = i]\!]$. The identity map $\mathrm{Id}[X] : X \to \Delta(X)$ is given by $x \mapsto \delta[x]$.

Graphically, the identity map is a bare line:

$$\mathrm{Id}[X] := \tag{8}$$

The erase map $*[A] : A \to \{1\}$ is the map $*[A]_i = 1$. It is the unique Markov kernel with domain A and only one column.

Graphically, the stopper is a fuse:

$$\mathrm{Id}[X] := \tag{9}$$

The copy map $\curlyvee[X] : X \to \Delta(X \times X)$ is the Markov kernel defined by $\curlyvee_x := \delta_x \otimes \delta_x$. Graphically it is a fork with a dot at the point where it splits:

$$\curlyvee[X] := \tag{10}$$

2.4 Products

Two Markov kernels $\mathbf{L} : X \to \Delta(Y)$ and $\mathbf{M} : Y \to \Delta(Z)$ have a product $\mathbf{LM} : X \to \Delta(Z)$ given by the usual matrix-matrix product: $\mathbf{LM}^z_x = \sum_y \mathbf{L}^y_x \mathbf{M}^z_y$. Graphically, we write represent products by joining kernel wires together:

$$\mathbf{LM} := \tag{11}$$

2.5 Label dictionaries, equality and inclusion

A *modelling context* $\mathcal{M}$ is a collection of labels, a label dictionary (defined in this section) and a collection of conditional probabilities (defined in section 2.6.

It can be thought of as a namespace that ensures that the properties defined in section 2.8 hold.

, if we respect the naming rules (section 2.8), any two conditional probabilities with the same name will be the same and any model containing two instances of the label name will assign probability 0 to any point at which the two labels differ.

A label collection is a collection of labels, each of which is assigned to a measurable set. For example, $\{\mathsf{X} : X, \mathsf{Y} : Y, \mathsf{X}_1 : X_1, \mathsf{X}_2 : X_2\}$. A sequence of labels is associated with the cartesian product of the sets associated with the individual labels. For example, $(\mathsf{X}, \mathsf{Y}) : X \times Y$.

A label dictionary is a set of statements of equality between labels and between labels and sequences of labels. For example, we might have the set $\{\mathsf{X} = (\mathsf{X}_1, \mathsf{X}_2), \mathsf{W} = \mathsf{X}\}$. The sets associated with the labels on each side of all equality statements must be the same.

If an equality $\mathsf{W} = \mathsf{X}$ appears in the label dictionary then any true expression involving conditional probabilities or labels remains true when all instances of X are replaced by W or vise-versa. A label X_1 is contained in X iff it is possible to derive from the label dictionary a statement of equality between X and a sequence of labels containing X_1.

Because it saves a lot of space, we will generally hold to the convention that a label X is associated with the set X. However, this convention sometimes fails, for example when we have two labels X_1 and X_2 that are associated with the same set – in such cases, we will explicitly define the relationship.

The trivial label $*$ always corresponds to the 1-element set $\{*\}$. Because $\{*\} \times A$ is isomorphic to A for any A, we can consider any label sequence to be isomorphic to the same label sequence with any number of copies of the trivial label appended. A sequence of labels that consists entirely of trivial labels is equivalent to the trivial label. An empty sequence is equivalent to the trivial label.

2.6 Labeled Markov kernels, conditional probabilities

A labeled Markov kernel $(\mathbf{K}, \mathsf{A}_C, \mathsf{B}_D)$ is a Markov kernel $\mathbf{K} : X \to \Delta(Y)$ along with a sequence of *domain labels* $\mathsf{A}_C := (\mathsf{A}_i)_{i \in C}$ and *codomain labels* $\mathsf{B}_D := (\mathsf{B}_i)_{i \in D}$. A label assignment is valid if the Markov kernel's domain matches the sets associated with its domain labels and its codomain is the same as the sets associated with its codomain labels.

A labeled probability distribution $\mathbb{P} \in \Delta(Y)$ comes with a sequence of codomain labels $(\mathsf{B}_i)_{i \in D}$ only.

A conditional probability $\mathbb{L}[\mathsf{A}_C | \mathsf{B}_D; \mathbf{K}]$ is a labeled kernel $(\mathbf{K}, \mathsf{A}_C, \mathsf{B}_D)$ along with an *ambient model* (Definition 2.6) $\mathbb{L}$.

Graphically, we place the labels on the wires of a conditional probability and include the ambient model in the name of the Markov kernel:

$$\mathbb{L}[\mathsf{B}_1\mathsf{B}_2|\mathsf{A}_1\mathsf{A}_2;\mathbf{K}] := \begin{matrix}\mathsf{A}_1\\ \mathsf{A}_2\end{matrix}\boxed{\mathbf{K}[\mathbb{L}]}\begin{matrix}\mathsf{B}_1\\ \mathsf{B}_2\end{matrix} \tag{12}$$

A sequence of labels is itself a label, so we can also bundle wires and their corresponding labels together:

$$\mathbb{L}[\mathsf{B}_1\mathsf{B}_2|\mathsf{A}_1\mathsf{A}_2;\mathbf{K}] = (\mathsf{A}_1,\mathsf{A}_2)\text{-}\boxed{\mathbf{K}[\mathbb{L}]}\text{-}(\mathsf{B}_1,\mathsf{B}_2) \tag{13}$$

If two conditional probabilities $\mathbb{L}[\mathsf{A}_C|\mathsf{B}_D;\mathbf{K}]$ and $\mathbb{M}[\mathsf{A}_C|\mathsf{B}_D;\mathbf{K}]$ share the same kernel, we will say $\mathbb{L}[\mathsf{A}_C|\mathsf{B}_D;\mathbf{K}] \overset{krn}{=} \mathbb{K}[\mathsf{A}_C|\mathsf{B}_D;\mathbf{K}]$.

2.7 Ambient model

An ambient model is a tuple $\mathbb{L} := (\mathbf{L}, D, E, \mathcal{C}, \mathcal{R})$ where $\mathbf{L} : D \to \Delta(E)$ is a Markov kernel, D and E are the domain and codomain, $\mathcal{C}$ is a set of *choice variable definitions* and $\mathcal{R}$ is a set of *random variable definitions*. $\mathcal{C}$ and $\mathcal{R}$ each assign labels to measurable functions on D and E respectively. Given a label X associated with a set X, if the label assignment $\mathsf{X} : f \in \mathcal{R}$ then it must be the case that $f : D \to X$ or $f : E \to X$.

If the label dictionary contains the assignment $\mathsf{X} = (\mathsf{Y}, \mathsf{Z})$ and the variable assignments $\{\mathsf{X} : f_{\mathsf{X}}, \mathsf{Y} : f_{\mathsf{Y}}, \mathsf{Z} : f_{\mathsf{Z}}\}$ are a subset of either $\mathcal{R}$ or $\mathcal{C}$, then it must be the case that $f_{\mathsf{X}}(h) = (f_{\mathsf{Y}}(h), f_{\mathsf{Z}}(h))$.

Let D be the sequence of all choice variables $\mathsf{D} = (\mathsf{A})_{\mathsf{A}\in\mathcal{D}}$. Then f_{D} must be an invertible function of D.

The same label can appear in both $\mathcal{C}$ and $\mathcal{R}$ and can appear multiple times in $\mathcal{R}$. However, each label can appear at most once in $\mathcal{C}$. In addition, the assignment of repeated labels to functions must be compatible with the first axiom of modelling contexts (2.8); namely, the random variables that are given the same name must be almost surely equal.

Given $\mathsf{X} : f_{\mathsf{X}} \in \mathcal{R}$, $\mathbb{L}[\mathsf{X}|\mathsf{D}] := (d, A) \mapsto \sum_{i\in f_{\mathsf{X}}^{-1}(A)} \mathbf{L}_d^i$ is the *conditional probability of* X *given* D under $\mathbb{L}$.

Given $\{\mathsf{X} : f_{\mathsf{X}}, \mathsf{Y} : f_{\mathsf{Y}}\} \in \mathcal{R}$, any Markov kernel $\mathbf{M}$ such that $\mathbb{L}[\mathsf{XY}|\mathsf{D}]_d^{A\times B} = \sum_{i\in f_{\mathsf{X}}^{-1}(A)} \mathbb{L}[\mathsf{X}|\mathsf{D}]_d^i \mathbf{M}(d,i)^B$ is a version of the *conditional probability of* Y *given* (X,D) under $\mathbb{L}$.

2.8 Modelling context axioms

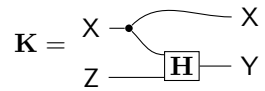
We place the following requirements on elements of $\mathcal{M}$:

1. Any model containing two instances of the same label will assign probability 0 to any point at which the two labels differ
2. The kernel $\mathbf{K} : A^C \to \Delta(B^D)$ of the conditional probability $\mathbb{L}[\mathsf{A}_C|\mathsf{B}_D;\mathbf{K}]$ is a version of the *probability of* A_C *given* B_D under the ambient model $\mathbb{L}$

The first axiom can be broken down into two cases.

Case 1 Given $\mathbb{M}[\mathsf{XY}|\mathsf{XZ};\mathbf{K}]$, we require that there exist some $\mathbf{H} : X \times Z \to \Delta(Y)$ such that

$$\mathbf{K} = \tag{14}$$

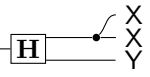


$$\iff \tag{15}$$

$$\mathbf{K}^{x'y}_{xz} = \delta[x]^{x'}\mathbf{L}^{y}_{xz} \tag{16}$$

Case 2 Given $\mathbb{M}[\mathsf{XXY}|\mathsf{Z};\mathbf{K}]$, we require that there exist some $\mathbf{H} : Z \to \Delta(X \times Y)$ such that

$$\mathbf{K} = \tag{17}$$



$$\iff \tag{18}$$

$$\mathbf{K}^{xx'y}_{z} = \delta[x]^{x'}\mathbf{L}^{xy}_{z} \tag{19}$$

2.9 Connection

Connection is an operation $\rightrightarrows$ that "joins" two labeled Markov kernels where the labels can be matched and preserves unmatched inputs and outputs. A key property of extension is that, if both input Markov kernels satisfy Axiom 1, then the output also satisfies axiom 1. Depending on the labels of the inputs, the extension operation can reduce to:

- The tensor product
- The matrix product
- The operation that combines a marginal and a conditional probability to yield a joint probability

We overload $\rightrightarrows$ to also refer to an conditional probabilities. In general, the result of connecting of two conditional probabilities sharing an ambient model does not satisfy Axiom 2. We assume the result of connecting two conditional probabilities is equipped with an arbitrary ambient model satisfying Axiom 2. We show that it is always possible to make some choice for this and, in particular cases, an ambient model matching the inputs can be chosen.

Given two labeled Markov kernels $(\mathbf{K}, \mathsf{I}_F, \mathsf{O}_F)$ and $(\mathbf{L}, \mathsf{I}_S, \mathsf{O}_S)$, make the

following label identifications:

$$\mathsf{O}_{F\cdot} := \mathsf{O}_F \setminus \mathsf{I}_S \tag{20}$$
$$\mathsf{O}_{FS} := \mathsf{O}_F \cap \mathsf{I}_S \tag{21}$$
$$\mathsf{I}_{F\cdot} := \mathsf{I}_F \setminus \mathsf{I}_S \tag{22}$$
$$\mathsf{I}_{FS} := \mathsf{I}_F \cap \mathsf{I}_S \tag{23}$$
$$\mathsf{I}_{\cdot S} := \mathsf{I}_S \setminus \mathsf{I}_F \tag{24}$$
$$\mathsf{O}_{I_F O_{S^*}} := \mathsf{O}_S \cap \mathsf{I}_F \tag{25}$$
$$\mathsf{O}_{O_F O_{S^*}} := \mathsf{O}_F \cap \mathsf{O}_S \setminus \mathsf{I}_S \tag{26}$$
$$\tag{27}$$

The output labels may contain duplicates, but not the input labels. We use the convention that at most one of each label in O_F belongs to O_{FS} and the multiplicity of each label in $(\mathsf{O}_{F\cdot}, \mathsf{O}_{FS}, \mathsf{O}_{\cdot S})$ is equal to the multiplicity of each label in $(\mathsf{O}_F, \mathsf{O}_S)$. Thus if a label is shared between O_F and I_S and appears 3 times in O_F, one copy is assigned to O_{FS} and two to $\mathsf{O}_{F\cdot}$.

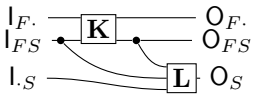
$(\mathbf{K}, \mathsf{I}_F, \mathsf{O}_F)$ can be connected to $(\mathbf{L}, \mathsf{I}_S, \mathsf{O}_S)$ iff $\mathsf{O}_{I_F O_{S^*}}$ is trivial (the output of the "second" kernel cannot be connected to the inputs of the "first") and $\mathsf{O}_{O_F O_{S^*}}$ is also trivial (the two kernels do not propose conflicting models of the same label).

Definition 2.1 (extension)**.** Consider two labeled Markov kernels $(\mathbf{K}, \mathsf{I}_F, \mathsf{O}_F)$ which can be connected to $(\mathbf{L}, \mathsf{I}_S, \mathsf{O}_S)$. Because they can be conected, we can write $(\mathbf{K}, (\mathsf{I}_{F\cdot}, \mathsf{I}_{FS}), (\mathsf{O}_{F\cdot}, \mathsf{O}_{FS}))$ and $(\mathbf{L}, (\mathsf{I}_{FS}, \mathsf{I}_{\cdot S}), \mathsf{O}_S)$.

Then Equations 29 and 30 are equivalent definitions of extension:

$$(\mathbf{K}, (\mathsf{I}_{F\cdot}, \mathsf{I}_{FS}), (\mathsf{O}_{F\cdot}, \mathsf{O}_{FS})) \rightrightarrows (\mathbf{L}, (\mathsf{I}_{FS}, \mathsf{I}_{\cdot S}), \mathsf{O}_S) := (\mathbf{J}, (\mathsf{I}_{F\cdot}, \mathsf{I}_{FS}, \mathsf{I}_{\cdot S}), (\mathsf{O}_{F\cdot}, \mathsf{O}_{FS}, \mathsf{O}_S)) \tag{28}$$

$:=$

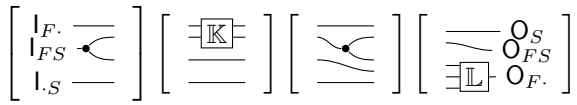


$$\tag{29}$$

$$\mathbf{J}^{zxw}_{yqr} = \mathbf{K}^{zx}_{yq}\mathbf{L}^{w}_{xqr} \tag{30}$$

Equation 29 can be broken down to the product of four Markov kernels, each of which is itself a tensor product of a number of other Markov kernels:

$(\mathbf{J}, (\mathsf{I}_{F\cdot}, \mathsf{I}_{FS}, \mathsf{I}_{\cdot S}), (\mathsf{O}_{F\cdot}, \mathsf{O}_{FS}, \mathsf{O}_S)) =$



$$\tag{31}$$

$$\tag{32}$$

Lemma 2.2 (Extension is associative up to permuation of labels)**.** *Given labeled Markov kernels* $(\mathbf{K}, \mathsf{I}_K, \mathsf{O}_K)$*,* $(\mathbf{L}, \mathsf{I}_L, \mathsf{O}_L)$ *and* $(\mathbf{J}, \mathsf{I}_J, \mathsf{O}_J)$*,*

$$(\mathbf{K} \rightrightarrows \mathbf{L}) \rightrightarrows \mathbf{J} \overset{perm}{=} \mathbf{K} \rightrightarrows (\mathbf{L} \rightrightarrows \mathbf{J}) \tag{33}$$

Where $\overset{perm}{=}$ *indicates equality up to permutation of labels and corresponding swaps applied to the Markov kernel.*

Proof. This will be proven with string diagrams, and consequently generalises to the operation defined by Equation 17 in other Markov kernel categories.

Define

$$\mathsf{I}_{K\cdot\cdot} := \mathsf{I}_K \setminus \mathsf{I}_L \setminus \mathsf{I}_J \tag{34}$$
$$\mathsf{I}_{KL\cdot} := \mathsf{I}_K \cap \mathsf{I}_L \setminus \mathsf{I}_J \tag{35}$$
$$\mathsf{I}_{K\cdot J} := \mathsf{I}_K \cap \mathsf{I}_J \setminus \mathsf{I}_L \tag{36}$$
$$\mathsf{I}_{KLJ} := \mathsf{I}_K \cap \mathsf{I}_L \cap \mathsf{I}_J \tag{37}$$
$$\mathsf{I}_{\cdot L\cdot} := \mathsf{I}_L \setminus \mathsf{I}_K \setminus \mathsf{I}_J \tag{38}$$
$$\mathsf{I}_{\cdot LJ} := \mathsf{I}_L \cap \mathsf{I}_J \setminus \mathsf{I}_K \tag{39}$$
$$\mathsf{I}_{\cdot\cdot J} := \mathsf{I}_J \setminus \mathsf{I}_K \setminus \mathsf{I}_L \tag{40}$$
$$\mathsf{O}_{K\cdot\cdot} := \mathsf{O}_K \setminus \mathsf{I}_N \setminus \mathsf{I}_J \tag{41}$$
$$\mathsf{O}_{KL\cdot} := \mathsf{O}_K \cap \mathsf{I}_L \setminus \mathsf{I}_J \tag{42}$$
$$\mathsf{O}_{K\cdot J} := \mathsf{O}_K \cap \mathsf{I}_J \setminus \mathsf{I}_L \tag{43}$$
$$\mathsf{O}_{KLJ} := \mathsf{O}_K \cap \mathsf{I}_L \cap \mathsf{I}_J \tag{44}$$
$$\mathsf{O}_{L\cdot} := \mathsf{O}_L \setminus \mathsf{I}_J \tag{45}$$
$$\mathsf{O}_{LJ} := \mathsf{O}_L \cap \mathsf{I}_J \tag{46}$$

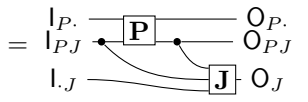
Also define

$$(\mathbf{P}, \mathsf{I}_P, \mathsf{O}_P) := \mathbf{K} \rightrightarrows \mathbf{L} \tag{47}$$
$$(\mathbf{Q}, \mathsf{I}_Q, \mathsf{O}_Q) := \mathbf{L} \rightrightarrows \mathbf{J} \tag{48}$$

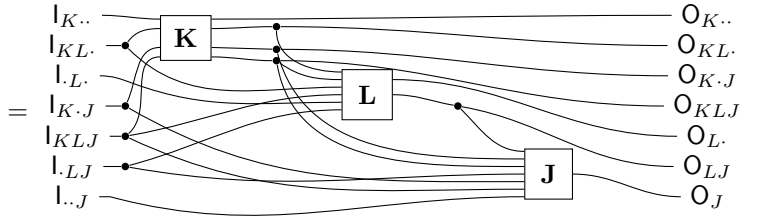
Then

$$(\mathbf{K} \rightrightarrows \mathbf{L}) \rightrightarrows \mathbf{J} = \mathbf{P} \rightrightarrows \mathbf{J} \tag{49}$$

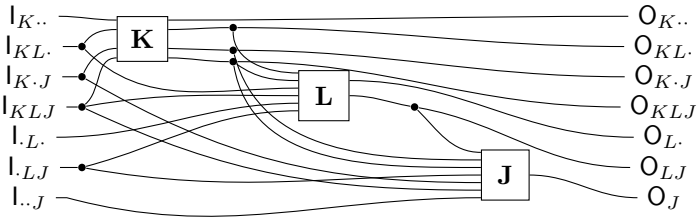
$$= \tag{50}$$



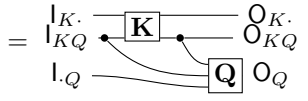
$$= \tag{51}$$



$$\overset{perm}{=} \tag{52}$$



$$= \tag{53}$$



□

Theorem 2.3 (Extension is compatible with a modelling context)**.** *Given* $\mathbb{M}[\mathsf{ZX}|\mathsf{YQ};\mathbf{K}]$ *and* $\mathbb{N}[\mathsf{W}|\mathsf{XQR};\mathbf{L}]$ *in a modelling context* $\mathcal{M}$*, let* $\mathbf{J} = \mathbb{M}[\mathsf{ZX}|\mathsf{YQ};\mathbf{K}] \rightrightarrows \mathbb{N}[\mathsf{W}|\mathsf{XQR};\mathbf{L}]$*. Then there exists an ambient model* $\mathbb{O}$ *such that* $\mathcal{M}$ *along with* $\mathbb{O}[\mathsf{ZXW}|\mathsf{YQR};\mathbf{J}]$ *is a valid modelling context, where we can choose* $\mathbb{O} = \mathbb{M}$ *if* $\mathbb{M} = \mathbb{N}$ *and the non-shared inputs* Y *and* R *are trivial.*

Proof. By inspecting the label set in $\mathbb{O}[\mathsf{ZXW}|\mathsf{YQR};\mathbf{J}]$, we can see that no labels from either of the inputs are duplicated. We need to verify that if either of the inputs had a duplicated label, then the result of the extension still satisfies axiom 1.

We have a number of cases to deal with Introduce $\mathbb{I}[\mathsf{X}|\mathsf{X};\mathrm{Id}_X]$ and note that in the case of Equation 14 corresponds to the equation

$$\mathbf{K} = \mathbb{I}[\mathsf{X}|\mathsf{X}; \mathrm{Id}_X] \rightrightarrows \mathbb{M}[\mathsf{Y}|\mathsf{XZ}; \mathbf{L}] \tag{54}$$

$$\implies \mathbf{J} = (\mathbb{I}[\mathsf{X}|\mathsf{X}; \mathrm{Id}_X] \rightrightarrows \mathbb{M}[\mathsf{Y}|\mathsf{XZ}; \mathbf{L}]) \rightrightarrows \mathbb{N}[\mathsf{W}|\mathsf{XQR}; \mathbf{L}] \tag{55}$$

$$\implies \mathbf{J} = \mathbb{I}[\mathsf{X}|\mathsf{X}; \mathrm{Id}_X] \rightrightarrows (\mathbb{M}[\mathsf{Y}|\mathsf{XZ}; \mathbf{L}] \rightrightarrows \mathbb{N}[\mathsf{W}|\mathsf{XQR}; \mathbf{L}]) \tag{56}$$

We can always define a new probability space so that $\mathbb{O}[\mathsf{ZXW}|\mathsf{YQR}; \mathbf{J}]$ satisfies axiom 2. We will show that if $\mathbb{M} = \mathbb{N}$ and the non-shared inputs Y and R are trivial then in particular we can choose $\mathbb{O} = \mathbb{M}$. □

Definition 2.4 (marginal)**.** Given a conditional probability $\mathbb{K}[\mathsf{XY}|\mathsf{W}]$, the marginal $\mathbb{K}[\mathsf{X}|\mathsf{W}]$ is defined as

$$\mathbb{K}[\mathsf{X}|\mathsf{W}] := \quad \mathsf{W} \; \boxed{\mathbb{K}} \; \mathsf{X} \tag{57}$$

$$\mathbb{K}[\mathsf{X}|\mathsf{W}]^x_w = \sum_{y \in Y} \mathbb{K}[\mathsf{XY}|\mathsf{W}]^{xy}_w \tag{58}$$

Definition 2.5 (disintegration)**.** $\mathbb{K}[\mathsf{Y}|\mathsf{XW}]$ is a disintegration of $\mathbb{K}[\mathsf{XY}|\mathsf{W}]$ if

$$\mathbb{K}[\mathsf{X}|\mathsf{W}] \rightrightarrows \mathbb{K}[\mathsf{Y}|\mathsf{XW}] = \mathbb{K}[\mathsf{XY}|\mathsf{W}] \tag{59}$$

Any Markov kernel $\mathbf{L}$ with the property

$$\mathbf{L}^y_{xw} = \frac{\mathbb{K}[\mathsf{XY}|\mathsf{W}]^{xy}_w}{\sum_{x \in X} \mathbb{K}[\mathsf{XY}|\mathsf{W}]^{xy}_w} \qquad \forall w, y: \text{ the denominator is positive} \tag{60}$$

is a version of $\mathbb{K}[\mathsf{Y}|\mathsf{XW}]$.

Definition 2.6 (ambient conditional probability)**.** A conditional probability $\mathbb{K}[\mathsf{Y}|\mathsf{X}]$ is an *ambient conditional probability* relative to $\mathcal{M}$ if there is no other conditional probability in $\mathcal{M}$ such that $\mathbb{K}[\mathsf{Y}|\mathsf{X}]$ is either a marginal or a disintegration of this conditional probability.

Recall that, if $\mathbb{K}[\mathsf{Y}|\mathsf{X}]$ is an ambient conditional probability, then $\mathbf{K}[\mathsf{Y}|\mathsf{X}] = \mathbf{K}$.

2.10 Conditional independence

Given $\mathbb{K}[\mathsf{X}|\mathsf{WZ}]$ in general we have no definition of $\mathbb{K}[\mathsf{X}|\mathsf{Z}]$. However, we can define such a "conditional probability" if we have the additional fact that X is independent of W given Z relative to $\mathbb{K}$.

Given $\mathbb{K}[\mathsf{X}|\mathsf{WZ}]$ we say X is independent of W given Z relative to $\mathbb{K}$, notated $\mathsf{X} \perp\!\!\!\perp_{\mathbb{K}} \mathsf{W}|\mathsf{Z}$ iff $\mathbb{K}[\mathsf{X}|\mathsf{WZ}]^x_{wz} = \mathbb{K}[\mathsf{X}|\mathsf{WZ}]^x_{w'z}$ for all $w, w' \in W$, $x \in X$ and $z \in Z$.

Given $\mathbb{K}[\mathsf{X}|\mathsf{WZ}]$ such that $\mathsf{X} \perp\!\!\!\perp_{\mathbb{K}} \mathsf{W}|\mathsf{Z}$, we define $\mathbb{K}[\mathsf{X}|\mathsf{Z}]$ to be any kernel satisfying $\mathbb{K}[\mathsf{X}|\mathsf{Z}]^x_z = \mathbb{K}[\mathsf{X}|\mathsf{DZ}]^x_{dz}$ for all x, z, d.

2.11 Uniqueness of disintegrations

Every conditional probability $\mathbf{K}[\mathsf{X}|\mathsf{Y}]$ is unique up to an equivalence class defined with respect to the ambient conditional probability $\mathbb{K}$.

Proof sketch: if it is an ambient conditional probability, then it is unique. If not, it is obtained from an ambient conditional probability by a sequence of marginalisations and disintegrations. Defining the equivalence class to be "equal up to measure 0 sets", marginalisations and disintegrations are both unique.

2.12 Existence of modelling context

Take a collection of Markov kernels and give them label sets consistent with their type signatures and respecting the rule that identical labels require identical spaces. Add all the recursive disintegrations + marginals. Add all valid extensions assigning a new model name for any result not already in the modelling context. Add all recursive disintegrations and marginals of valid extensions, etc.

Then: disintegration, marginalisation, extension operations all preserve label consistency rules. By construction, marginals, disintegrations and extensions are included. Marginalisation + disintegration preserves uniqueness of ambient conditional probability. Extension + assigning a new model name also preserves uniqueness of ambient conditional probability.

2.13 Standard probability models

The operation of combining two conditional probabilities which do not share a model name and obtaining a conditional probability relative with a new model name is unique to our approach. The standard approach to probability modelling features an ambient probability distribution defined on a sample space, along with "labels" that each correspond to a measurable functions on the sample space. With this setup, we can define marginals and disintegrations with respect to any sequences of labels. It is an open question whether there is a way to construct a modelling context with a single model that is equivalent to a standard probability model.

3 See-do models

Modular probability is useful when we want to combine different Markov kernels in such a way that "variables" refer to something consistent even though they don't necessarily have a unique distribution. The first example we will present is using modular probability to model decision problems.

Suppose we will be given an observation $x \in X$ and in response to this we can select any decision or stochastic mixture of decisions from a set D; that is we can choose a "strategy" as any Markov kernel $\mathbf{S}_\alpha : X \to \Delta(D)$. We are intersted in forecasting some consequences that take values in some set Y, and comparing the forecasts for different strategy choices so as to choose a best strategy.

How can we model this? One way to proceed is as follows: Define a model context $\mathcal{M}$ to which we add the conditional probabilities mentioned hereafter. For each strategy $\mathbb{S}_\alpha[\mathsf{D}|\mathsf{X}]$, our forecast will be represented by some joint probability in $\mathbb{P}_\alpha[\mathsf{XDY}|\mathsf{H}]$ where H is associated with a set of hypotheses H representing different choices that we think might be reasonable to make that may lead to different forecasts. Because observations come before we execute our strategy, we assume that $\mathbb{P}_\alpha[\mathsf{X}|\mathsf{H}] = P_\beta[\mathsf{X}|\mathsf{H}]$ for all α, β. Our chosen strategy is the probability of D given X: $\mathbb{P}_\alpha[\mathsf{D}|\mathsf{X}] \overset{krn}{=} \mathbb{S}_\alpha[\mathsf{D}|\mathsf{X}]$. Finally, our forecast of Y is the same for all strategies holding the observations, the decision and the hypothesis fixed: $\mathbb{P}_\alpha[\mathsf{Y}|\mathsf{HD}] = P_\beta[\mathsf{Y}|\mathsf{HD}]$ for all α, β.

Under these assumptions, there exists $\mathbb{T}[\mathsf{XY}|\mathsf{HD}] \in \mathcal{M}$ with $\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$ such that for all α,

$$\mathbb{P}_\alpha[\mathsf{XDY}|\mathsf{H}] \overset{krn}{=} \mathbb{T}[\mathsf{X}|\mathsf{H}] \rightrightarrows \mathbb{S}_\alpha[\mathsf{D}|\mathsf{X}] \rightrightarrows \mathbb{T}[\mathsf{Y}|\mathsf{XHD}] \tag{61}$$

The proof is given in Appendix 6. Note that $\mathbb{T}[\mathsf{X}|\mathsf{H}]$ exists by virtue of the fact $\mathsf{X} \perp\!\!\!\perp_{\mathbb{T}} \mathsf{D}|\mathsf{H}$. While this independence is what enables Equation 61, in general $\mathsf{X} \not\perp\!\!\!\perp_{\mathbb{P}_\alpha} \mathsf{D}|\mathsf{H}$, so $\mathbb{T}$ cannot be a disintegration of $\mathbb{P}_\alpha$. Modular probability allows us to specify $\mathbb{T}$, which we call a *see-do model*, as a partial forecast to be completed with a strategy $\mathbb{S}_\alpha$ while also being able to use consistent names for variables that represent the same things (observations, decisions, consequences, hypotheses) whether their distributions are given by $\mathbb{P}_\alpha$, $\mathbb{T}$, which are mutually incompatible conditional probabilities.

3.1 See-do models and classical statistics

A *statistical model* (or *statistical experiment*) is a collection of probability distributions indexed by some set Θ. We can observe that $\{\mathbb{T}[\mathsf{X}|\mathsf{H}]_h\}_{h \in H}$ is a collection of probability distributions indexed by H.

In statistical decision theory, as introduced by Wald (1950), we are given a statistical experiment $\{\mathbb{P}_\theta \in \Delta(X)\}_\Theta$, a decision set D and a loss $l : \Theta \times D \to \mathbb{R}$. A strategy $\mathbb{S}_\alpha : X \to \Delta(D)$ is evaluated according to the risk functional $R(\theta, \mathbb{S}_\alpha) = \sum_{x \in X} \sum_{d \in D} \mathbb{P}_\theta^x (S_\alpha)_x^d l(h, d)$.

Suppose we have a see-do model $\mathbb{T}[\mathsf{XY}|\mathsf{HD}]$ with $\mathsf{Y} \perp\!\!\!\perp_{\mathbf{T}} \mathsf{X}|\mathsf{HD}$, and suppose that the random variable Y is a "reverse utility" function taking values in $\mathbb{R}$ for which low values are considered desirable. Then, defining a loss $l : H \times D \to \mathbb{R}$ by $l(h, d) = \sum_{y \in \mathbb{R}} y \mathbb{T}[\mathsf{Y}|\mathsf{HD}]_{h,d}^y$, we have

$$\mathbb{E}_{\mathbb{P}_\alpha[\mathsf{XDY|H}]}[\mathsf{Y}] = \sum_{x\in X}\sum_{d\in D}\sum_{y\in Y} y\,(\mathbb{T}[\mathsf{X|H}] \rightrightarrows \mathbb{S}_\alpha[\mathsf{D|X}] \rightrightarrows \mathbb{T}[\mathsf{Y|XHD}])_h^{xdy} \tag{62}$$

$$= \sum_{x\in X}\sum_{d\in D}\sum_{y\in Y} \mathbb{T}[\mathsf{X|H}]_h^x \mathbb{S}_\alpha[\mathsf{D|X}]_x^d \mathbb{T}[\mathsf{Y|HD}]_{h,d}^y \tag{63}$$

$$= \sum_{x\in X}\sum_{d\in D} \mathbb{T}[\mathsf{X|H}]_h^x (S_\alpha)_x^d l(h,d) \tag{64}$$

$$= R(h, \mathbb{S}_\alpha) \tag{65}$$

That is, if we are given a see-do model where we interpret $\mathbb{T}[\mathsf{X|H}]$ as a statistical experiment and Y as a reversed utility, the expectation of the utility under the strategy forecast given in equation 61 is the risk of that strategy under hypothesis h.

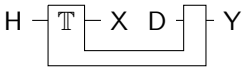
3.2 Combs

The see-do model $\mathbb{T}[\mathsf{XY|HD}]$ is known as a *comb*. This structure was introduced by Chiribella et al. (2008) in the context of quantum circuit architecture, and Jacobs et al. (2019) adapted the concept to causal modelling.

A comb is a Markov kernel with a "hole" in it. We combine the see-do model with a strategy by putting the strategy "in the middle" of the see-do model (Equation 61), rather than attaching it to one end. While it is not a well-formed diagram in the language described in this paper, we can visualise combs as Markov kernels with holes:

$\mathbb{T}[\mathsf{XY|HD}] =$ 

(66)

$=$ 

(67)

We can take any strategy $\mathbb{S}_\alpha[\mathsf{D|X}]$ and drop it into the "hole" in 67 (as described in Equation 61) to get a forecast of the outcome of that strategy.

4 Causal Bayesian Networks

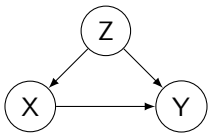
A Causal Bayesian Network posits a set of observational probability distributions, for example $\{\mathbb{P}_h(X,Y)\}_H$, and a set of interventional distributions, for example $\{\mathbb{P}_h(X,Y|do(X=x))\}_{x\in X, h\in H}$. Here we use notation similar to typical notation used for Causal Bayesian Networks and don't intend for these to necessarily be elements of any modelling context. For simplicity, we will consider a Causal Bayesian Network with only hard interventions on a single variable, e.g. interventions only of the form $do(\mathsf{X}=x)$.

First we will offer some commentary

We can consider this an instance of a see-do model. To do so consistently within a modelling context $\mathcal{M}$, we need to distinguish observation and intervention variables - let the former retain the labels X, Y and call the latter X', Y'. Let $D = \{do(\mathsf{X} = x)\}_{x \in X}$. Then a Causal Bayesian Network can be considered a see-do model $\mathbb{T}[\mathsf{XYX'Y'|HD}]$ by identifying $\mathbb{T}[\mathsf{XY|H}]_h := \mathbb{P}_h(X, Y)$ and $\mathbb{T}[\mathsf{X'Y'|HD}]_{h,do(\mathsf{X}=x)} := P_h(X, Y|do(X = x))$.

We need to rename the consequence variables because otherwise we would have $\mathbb{T}[\mathsf{XYXY|HD}]$ and the two X's and the two Y's would be deterministically equal by the "identical labels" rule

We can say a bit more about Causal Bayesian Networks. Suppose we have the network



Then, letting $\mathbb{T}[\mathsf{XYZ|H}]$ be the observational "see" model and $\mathbb{T}[\mathsf{X'Y'Z'|HD}]$ be the interventional "do" model with D the set of interventions $\{do(\mathsf{X} = x)\}_{x \in X}$ where we write $x := do(\mathsf{X} = x)$ for short, then we know by the backdoor adjustment rule that $\mathbb{T}[\mathsf{X'Y'Z'|HD}]^{x'yz}_{hx} \overset{krn}{=} \mathbb{T}[\mathsf{Z|H}]^{z}_{h} \delta[x]^{x'} \mathbb{T}[\mathsf{Y|XZH}]^{y}_{hx'z}$.

Let $\mathbb{U}[\mathsf{ZY|XH}] = \mathbb{T}[\mathsf{Z|H}] \rightrightarrows \mathbb{T}[\mathsf{Y|XZH}]$, call $\mathbb{T}[\mathsf{X|H}]$ the "observational strategy" and $\mathbb{D}_x[\mathsf{X|D}]^{x'}_{x} \overset{krn}{=} \delta[x]^{x'}$ the interventional strategies for all $x \in X$. Then we have

$$\mathbb{T}[\mathsf{XYZ|H}] = \mathbb{U}[\mathsf{Z|H}] \rightrightarrows \mathbb{T}[\mathsf{X|H}] \rightrightarrows \mathbb{U}[\mathsf{Y|XHZ}] \tag{68}$$

$$\mathbb{T}[\mathsf{X'Y'Z'|HD}] \overset{krn}{=} \mathbb{U}[\mathsf{Z|H}] \rightrightarrows \mathbb{D}[\mathsf{X|D}] \rightrightarrows \mathbb{U}[\mathsf{Y|XHZ}] \tag{69}$$

So this simple example of a Causal Bayesian network is a "nested comb" where the outer comb $\mathbb{T}[\mathsf{XYZX'Y'Z'|HD}]$ is the "see" and "do" models, which are themselves generated by the inner comb $\mathbb{U}[\mathsf{ZY|XH}]$ with different choices $\mathbb{T}[\mathsf{X|H}]$ and $\mathbb{D}[\mathsf{X|D}]$ for the insert.

This is a simple example, but Jacobs et al. (2019) has used an "inner comb" representation of a general class of Causal Bayesian Networks to prove a sufficient identification condition which is itself slightly more general than the identification condition given by Tian and Pearl (2002).

5 Potential outcomes with and without counterfactuals

Potential outcomes is a widely used approach to causal modelling characterised by its use of "potential outcome" random variables. Potential outcome random variables are typically noted for being given counterfactual interpretations.

For example, suppose have something we want to model, call it TYT ("The Y Thing"), which we represent with a variable Y. Suppose we want to know how TYT behaves under different regimes 0 and 1 under which we want to know about TYT, and we use a variable W to indicate which regime holds at a given point in time. A potential outcomes model will introduce the two additional "potential outcome" variables $(\mathsf{Y}(0), \mathsf{Y}(1))$. What these variables represent can be given a counterfactual interpretation like "$\mathsf{Y}(0)$ represents what TYT would be under regime 0, whether or not regime 0 is the actual regime" and similarly "$\mathsf{Y}(1)$ represents what TYT would be under regime 1, whether or not regime 1 is the actual regime". Note that we say "what TYT would be" rather that "what Y would be" as "what would Y be if W was 0 if W was actually 1" is not a question we can ask of random variables, but it is one that might make sense for the things we use random variables to model.

This is a key point, so it is worth restating: the assumption that potential outcome variables agree with "the value TYT would take" under fixed regimes regardless of the "actual" value of the regime seems to be a critical assumption that distinguishes potential outcome variables from arbitrary random variables that happen to take values in the same space as Y. However, this assumption can only be stated by making reference to the informally defined "TYT" and the informal distinction between the supposed and the actual value of the regime.

The potential outcomes framework features other critical assumptions that relate potential outcome variables to things that are only informally defined. For example, Rubin (2005) defines the *Stable Unit Treatment Value Assumption* (SUTVA) as:

> SUTVA (stable unit treatment value assumption) [...] comprises two subassumptions. First, it assumes that there is no interference between units (Cox 1958); that is, neither $Y_i(1)$ nor $Y_i(0)$ is affected by what action any other unit received. Second, it assumes that there are no hidden versions of treatments; no matter how unit i received treatment 1, the outcome that would be observed would be $Y_i(1)$ and similarly for treatment 0

"Versions of treatments" do not appear within typical potential outcomes models, so this is also an assumption about how "the thing we are trying to model" behaves rather than an assumption stated within the model.

Given informal assumptions like this, one may be motivated to "formalize" them. More specifically, one might be motivated to ask whether there is some larger class of models that, under conditions corresponding to the informal conditions above yield regular potential outcome models?

I have a vague intuition here that you always need some kind of assumption like "my model is faithful to the real thing", but if you are stating fairly specific conditions in English you should also be able to state them mathematically. Among other reasons, this is useful because it's easier for other people to know what you mean when you state them.

The approach we have introduced here, motivated by decision problems, has in the past been considered a means of avoiding counterfactual statements, which has been considered a positive by some (Dawid, 2000) and a negative by others:

> [...] Dawid, in our opinion, incorrectly concludes that an approach to causal inference based on "decision analysis" and free of counterfactuals is completely satisfactory for addressing the problem of inference about the effects of causes.(Robins and Greenland, 2000)

It may be surprising to some, then, that we can use see-do models to formally state these key assumptions associated with potential outcomes models. Furthermore, we will argue that potential outcomes are typically a strategy to motivate inductive assumptions in see-do models, and we will show that the counterfactual interpretation is unnecessary for this purpose.

5.1 Potential outcomes in see-do models

A basic property of potential outcomes models is the relation between variables representing actual outcomes and variables representing potential outcomes, which was stated informally in the opening paragraph of this section.

In the following definition, $\mathsf{Y}(W) = (\mathsf{Y}(w))_{w \in W}$.

Definition 5.1 (Potential outcomes)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, a collection of variables $\{\mathsf{Y}, \mathsf{Y}(W), \mathsf{W}\}$ where Y and $\mathsf{Y}(W)$ are random variables and W could be either a state or a random variable is a *potential outcome submodel* if $\mathbf{K}[\mathsf{Y}|\mathsf{W}\mathsf{Y}(W)]$ exists and $\mathbf{K}[\mathsf{Y}|\mathsf{W}\mathsf{Y}(W)]_{ij_1j_2\dots j_{|W|}} = \delta[j_i]$.

How this will change: a potential outcomes model is a comb $\mathbb{K}[\mathsf{Y}(W)|\mathsf{H}] \rightrightarrows \mathbb{K}[\mathsf{Y}|\mathsf{W}\mathsf{Y}(W)]$.

We allow X to be a state or a random variable to cover the cases where potential outcomes models feature as submodels of observation models (in which case X is a random variable) or as submodels of consequence models (in which case X may be a state variable).

As an aside that we could define stochastic potential outcomes if we allow the variables $\mathsf{Y}(x)$ to take values in $\Delta(Y)$ rather than in Y, and then require $\mathbf{K}[\mathsf{Y}|\mathsf{X}\mathsf{Y}(X)]_{ij_1j_2\dots j_{|X|}} = j_i$ (where j_i is an element of $\Delta(Y)$). This is more complex to work with and rarely seen in practice, but it is worth noting that Definition 5.1 can be generalised to cover models where $\mathsf{Y}(x)$ describes the value Y would take if X were x *with uncertainty*.

An arbitrary see-do model featuring potential outcome submodels does not necessarily allow for the formal statement of the counterfactual interpretation of potential outcomes. Here we use TYT ("the actual thing") and "regime" to refer to the things we are actually trying to model. We require that $\mathsf{Y} \overset{a.s.}{=} \mathsf{Y}(w)$ conditioned on $\mathsf{W} = w$. If we add an interpretation to this model saying Y represents TYT and W represents the regime, then we have "for all w, $\mathsf{Y}(w)$ is

equal to Y which represents TYT under the regime w". However, this does not guarantee that our model has anything that reasonably represents "what TYT would be equal to under supposed regime w if the regime is actually w'".

We propose *parallel potential outcome submodels* as a means of formalising statements about what how TYT behaves under "supposed" and "actual" regimes:

Definition 5.2 (Parallel potential outcomes)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, a collection of variables $\{\mathsf{Y}_i, \mathsf{Y}(W), \mathsf{W}_i\}$, $i \in [n]$, where Y_i and $\mathsf{Y}(W)$ are random variables and W_i could be either a state or random variables is a *parallel potential outcome submodel* if $\mathbf{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]$ exists and $\mathbf{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]_{kj_1j_2\ldots j_{|W|}} = \delta[j_k]$.

How this will change: a parallel potential outcomes model is a comb $\mathbb{K}[\mathsf{Y}(W)|\mathsf{H}] \rightrightarrows \mathbb{K}[\mathsf{Y}_i|\mathsf{W}_i\mathsf{Y}(W)]$.

A parallel potential outcomes model features a sequence of n "parallel" outcome variables Y_i and n "regime proposals" W_i, with the property that if the regime proposal $\mathsf{W}_i = w_i$ then the corresponding outcome $\mathsf{Y}_i \overset{a.s.}{=} \mathsf{Y}(w_i)$. We can identify a particular index, say $n = 1$, with the actual world and the rest of the indices with supposed worlds. Thus Y_1 represents the value of TYT in the actual world and Y_i $i \neq 1$ represents TYT under a supposed regime W_i. Given such an interpretation, the fact that $\mathsf{Y}_i \overset{a.s.}{=} \mathsf{Y}(w_i)$ can be interpreted as assuming "for all w, if the supposed regime W_i is w then the corresponding outcome will be almost surely equal to $\mathsf{Y}(w)$, regardless of the value of the actual regime W_1", which is our original counterfactual assumption.

We do not intend to defend this as the only way that counterfactuals can be modeled, or even that it is appropriate to capture the idea of counterfactuals at all. It is simply a way that we can model the counterfactual assumption typically associated with potential outcomes. We will show show that parallel potential outcome submodels correspond precisely to *extendably exchangeable* and *deterministically reproducible* submodels of Markov kernel spaces.

5.2 Parallel potential outcomes representation theorem

Exchangeble sequences of random variables are sequences whose joint distribution is unchanged by permutation. Independent and identically distributed random variables are one example: if X_1 is the result of the first flip of a coin that we know to be fair and X_2 is the second flip then $\mathbb{P}[\mathsf{X}_1\mathsf{X}_2] = \mathbb{P}[\mathsf{X}_2\mathsf{X}_1]$. There are also many examples of exchangeable sequences that are not mutually independent and identically distributed – for example, if we want to use random variables Y_1 and Y_2 to model our subjective uncertainty regarding two flips of a coin of unknown fairness, we regard our initial uncertainty for each flip to be equal $\mathbb{P}[\mathsf{Y}_1] = \mathbb{P}[\mathsf{Y}_2]$ and we our state of knowledge of the second flip after observing only the first will be the same as our state of knowledge of the first flip after observing only the second $\mathbb{P}[\mathsf{Y}_2|\mathsf{Y}_1] = \mathbb{P}[\mathsf{Y}_1|\mathsf{Y}_2]$, then our model of subjective uncertainty is exchangeable.

De Finetti's representation theorem establishes the fact that any infinite exchangeable sequence $\mathsf{Y}_1, \mathsf{Y}_2, ...$ can be modeled by the product of a *prior* probability $\mathbb{P}[\mathsf{J}]$ with J taking values in the set of marginal probabilities $\Delta(Y)$ and a conditionally independent and identically distributed Markov kernel $\mathbb{P}[\mathsf{Y}_A|\mathsf{J}]^{y_A}_j = \prod_{i \in A} \mathbb{P}[\mathsf{Y}_1|\mathsf{J}]^{y_i}_j$.

We extend the idea of exchangeable sequences to cover both random variables and state variables, and we show that a similar representation theorem holds for potential outcomes. De Finetti's original theorem introduced the variable J that took values in the set of marginal distributions over a single observation; the set of potential outcome variables plays an analagous role taking values in the set of functions from propositions to outcomes.

The representation theorem for potential outcomes is somewhat simpler that De Finetti's original theorem due to the fact that potential outcomes are usually assumed to be *deterministically reproducible*; in the parallel potential outcomes model, this means that for $j \neq i$, if W_j and W_i are equal then Y_j and Y_i will be almost surely equal. This assumption of determinism means that we can avoid appeal to a law of large numbers in the proof of our theorem.

An interesting question is whether there is a similar representation theorem for potential outcomes without the assumption of deterministic reproducibility. I'm reasonably confident that this is a straightforward corollary of the representation theorem proved in my thesis. However, this requires maths not introduced in this draft of the paper.

Extendably exchangeable sequences can be permuted without changing their conditional probabilities, and can be extended to arbitrarily long sequences while maintaining this property. We consider here sequences that are exchangeable conditional on some variable; this corresponds to regular exchageability if the conditioning variable is $*$ where $*_i = 1$.

Definition 5.3 (Exchangeability)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i \in [n]}$ with Y_i random variables is *exchangeable* conditional on Z if, defining $\mathsf{Y}_{[n]} = (\mathsf{Y}_i)_{i \in [n]}$ and $\mathsf{D}_{[n]} = (\mathsf{D}_i)_{i \in [n]}$, $\mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$ exists and for any bijection $\pi : [n] \to [n]$ $\mathbf{K}[\mathsf{Y}_{\pi([n])}|\mathsf{D}_{\pi([n])}\mathsf{Z}] = \mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$.

Definition 5.4 (Extension)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, $(\mathbf{K}', E', F')$ is an *extension* of $(\mathbf{K}, E, F)$ if there is some random variable X and some state variable U such that $\mathbf{K}'[\mathsf{X}|\mathsf{U}]$ exists and $\mathbf{K}'[\mathsf{X}|\mathsf{U}] = \mathbf{K}$.

If $(\mathbf{K}', E', F')$ is an extension of $(\mathbf{K}, E, F)$ we can identify any random variable Y on $(\mathbf{K}, E, F)$ with $\mathsf{Y} \circ \mathsf{X}$ on $(\mathbf{K}', E', F')$ and any state variable D with $\mathsf{D} \circ \mathsf{U}$ on $(\mathbf{K}', E', F')$ and under this identification $\mathbf{K}'[\mathsf{Y} \circ \mathsf{X}|\mathsf{D} \circ \mathsf{E}]$ exists iff $\mathbf{K}[\mathsf{Y}|\mathsf{D}]$ exists and $\mathbf{K}'[\mathsf{Y} \circ \mathsf{X}|\mathsf{D} \circ \mathsf{E}] = \mathbf{K}[\mathsf{Y}|\mathsf{D}]$. To avoid proliferation of notation, if we propose $(\mathbf{K}, E, F)$ and later an extension $(\mathbf{K}', E', F')$, we will redefine $\mathbf{K} := \mathbf{K}'$ and $\mathsf{Y} := \mathsf{Y} \circ \mathsf{X}$ and $\mathsf{D} := \mathsf{D} \circ \mathsf{E}$.

I think this is a very standard thing to do – propose some X and $\mathbb{P}(\mathsf{X})$ then introduce some random variable Y and $\mathbb{P}(\mathsf{XY})$ as if the sample space contained both X and Y all along.

Definition 5.5 (Extendably exchangeable)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ and a state variable Z with Y_i random variables is *extendably exchangeable* if there exists an extension of $\mathbf{K}$ with respect to which $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in\mathbb{N}}$ is exchangeable conditional on Z.

Here that we identify Z and $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ defined on the extension with the original variables defined on $(\mathbf{K}, E, F)$ while $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in\mathbb{N}\setminus[n]}$ may be defined only on the extension.

Deterministically reproducible sequences have the property that repeating the same decision gets the same response with probability 1. This could be a model of an experiment that exhibits no variation in results (e.g. every time I put green paint on the page, the page appears green), or an assumption about collections of "what-ifs" (e.g. if I went for a walk an hour ago, just as I actually did, then I definitely would have stubbed my toe, just like I actually did). Incidentally, many consider that this assumption is false concering what-if questions about things that exhibit quantum behaviour.

Definition 5.6 (Deterministically reproducible)**.** Given a Markov kernel space $(\mathbf{K}, E, F)$, a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ with Y_i random variables is *deterministically reproducible* conditional on Z if $n \geq 2$, $\mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$ exists and $\mathbf{K}[\mathsf{Y}_{\{i,j\}}|\mathsf{D}_{\{i,j\}}\mathsf{Z}]^{lm}_{kk} = [\![l = m]\!]\mathbf{K}[\mathsf{Y}_i|\mathsf{D}_i\mathsf{Z}]^l_k$ for all i, j, k, l, m.

Theorem 5.7 (Potential outcomes representation)**.** *Given a Markov kernel space $(\mathbf{K}, E, F)$ along with a sequence of variables $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ with $n \geq 2$ and a conditioning variable Z, $(\mathbf{K}, E, F)$ can be extended with a set of variables $\mathsf{Y}(D) := (\mathsf{Y}(i))_{i\in D}$ such that $\{\mathsf{Y}_i, \mathsf{Y}(D), \mathsf{D}_i\}$ is a parallel potential outcome submodel if and only if $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ is extendably exchangeable and deterministically reproducible conditional on Z.*

Proof. If: Because $((\mathsf{D}_i, \mathsf{Y}_i))_{i\in[n]}$ is extendably exchangeable, we can without loss of generality assume $n \geq |D|$.

Let $e = (e_i)_{i\in[|D|]}$. Introduce the variable $\mathsf{Y}(i)$ for $i \in D$ such that $\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_{[D]}\mathsf{Z}]_{ez} = \mathbf{K}[\mathsf{Y}_D|\mathsf{D}_D\mathsf{Z}]_{ez}$ and introduce X_i, $i \in D$ such that $\mathbf{K}[\mathsf{X}_i|\mathsf{D}_i\mathsf{Z}\mathsf{Y}(D)]^{x_i}_{e_i z j_1 \ldots j_{|D|}} = \delta[j_{e_i}]^{x_i}$. Clearly $\{\mathsf{X}_{[n]}, \mathsf{D}_{[n]}, \mathsf{Y}(D)\}$ is a parallel potential outcome submodel. We aim to show that $\mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}] = \mathbf{K}[\mathsf{X}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]$.

Let $y := (y_i)_{i\in|D|} \in Y^{|D|}$, $d := (d_i)_{i\in[n]} \in D^{[n]}$, $x := (x_i)_{i\in[n]} \in Y^{[n]}$.

$$\mathbf{K}[\mathsf{X}_n|\mathsf{D}_n\mathsf{Z}]^x_{dz} = \sum_{y\in Y^{|D|}} \mathbf{K}[\mathsf{X}_{[n]}|\mathsf{D}_n\mathsf{Z}\mathsf{Y}(D)]^x_{dzy}\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_{[n]}\mathsf{Z}]^y_{dz} \tag{70}$$

$$= \sum_{y\in Y^{|D|}} \prod_{i\in[n]} \delta[y_{d_i}]^{x_i}\mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_n\mathsf{Z}]^y_{dz} \tag{71}$$

Wherever $d_i = d_j := \alpha$, every term in the above expression will contain the product $\delta[\alpha]^{x_i}\delta[\alpha]^{x_j}$. If $x_i \neq x_j$, this will always be zero. By deterministic reproducibility, $d_i = d_j$ and $x_i \neq x_j$ implies $\mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]_d z^x = 0$ also. We need to check for equality for sequences x and d such that wherever $d_i = d_j$,

$x_i = x_j$. In this case, $\delta[\alpha]^{x_i}\delta[\alpha]^{x_j} = \delta[\alpha]^{x_i}$. Let $Q_d \subset [n] := \{i| \nexists i \in [n] : j < i \ \& \ d_j = d_i\}$, i.e. Q is the set of all indices such that d_i is the first time this value appears in d. Note that Q_d is of size at most $|D|$. Let $Q_d^C = [n] \setminus Q_d$, let $R_d \subset D : \{d_i | i \in Q_d\}$ i.e. all the elements of D that appear at least once in the sequence d and let $R_d^C = D \setminus R_d$.

Let $y' = (y_i)_{i \in Q_d^C}$, $x_{Q_d} = (x_i)_{i \in Q_d}$, $\mathsf{Y}(R_d) = (\mathsf{Y}_d)_{d \in R_d}$ and $\mathsf{Y}(S_d) = (\mathsf{Y}_d)_{d \in S_d}$.

$$\mathbf{K}[\mathsf{X}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]^x_{dz} = \sum_{y \in Y^{|D|}} \prod_{i \in Q_d} \delta[y_{d_i}]^{x_i} \mathbf{K}[\mathsf{Y}(D)|\mathsf{D}_{[n]}\mathsf{Z}]^y_{dz} \tag{72}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}(R_d)\mathsf{Y}(R_d^C)|\mathsf{D}_{Q_d}\mathsf{D}_{Q_d^C}\mathsf{Z}]^{x_{Q_d}y'}_{d_{Q_d}d^C_{Q_d}z} \tag{73}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}_{R_d}\mathsf{Y}_{R_d^C}|\mathsf{D}_{Q_d}\mathsf{D}_{Q_d^C}\mathsf{Z}]^{x_{Q_d}y'}_{dz} \tag{74}$$

$$= \sum_{y' \in Y^{|R_d^C|}} \mathbf{K}[\mathsf{Y}_{[n]}|\mathsf{D}_{[n]}\mathsf{Z}]^{x_{Q_d}y'}_{dz} \qquad \text{(using exchangeability)} \tag{75}$$

Note that

Only if: We aim to show that the sequences $\mathsf{Y}_{[n]}$ and $\mathsf{D}_{[n]}$ in a parallel potential outcomes submodel are exchangeable and deterministically reproducible. □

6 Appendix:see-do model representation

Modularise the treatment of probability

Theorem 6.1 (See-do model representation)**.** *Suppose we have a decision problem that provides us with an observation $x \in X$, and in response to this we can select any decision or stochastic mixture of decisions from a set D; that is we can choose a "strategy" as any Markov kernel $\mathbf{S} : X \to \Delta(D)$. We have a utility function $u : Y \to \mathbb{R}$ that models preferences over the potential consequences of our choice. Furthermore, suppose that we maintain a denumerable set of hypotheses H, and under each hypothesis $h \in H$ we model the result of choosing some strategy $\mathbf{S}$ as a joint probability over observations, decisions and consequences $\mathbb{P}_{h,\mathbf{S}} \in \Delta(X \times D \times Y)$.*

Define X, Y and D such that $\mathsf{X}_{xdy} = x$, $\mathsf{Y}_{xdy} = y$ and $\mathsf{D}_{xdy} = d$. Then making the following additional assumptions:

1. *Holding the hypothesis h fixed the observations as have the same distribution under any strategy: $\mathbb{P}_{h,\mathbf{S}}[\mathsf{X}] = \mathbb{P}_{h,\mathbf{S}''}[\mathsf{X}]$ for all $h, \mathbf{S}, \mathbf{S}'$ (observations are given "before" our strategy has any effect)*

2. *The chosen strategy is a version of the conditional probability of decisions given observations: $\mathbf{S} = \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]$*

3. *There exists some strategy* $\mathbf{S}$ *that is strictly positive*

4. *For any* $h \in H$ *and any two strategies* $\mathbf{Q}$ *and* $\mathbf{S}$*, we can find versions of each disintegration such that* $\mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{DX}] = \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{DX}]$ *(our strategy tells us nothing about the consequences that we don't already know from the observations and decisions)*

Then there exists a unique see-do model $(\mathbf{T}, \mathsf{H}', \mathsf{D}', \mathsf{X}', \mathsf{Y}')$ *such that* $\mathbb{P}_{h,\mathbf{S}}[\mathsf{XDY}]^{ijk} = \mathbf{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbf{S}^j_i \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ijk}$.

Proof. Consider some probability $\mathbb{P} \in \Delta(X \times D \times Y)$. By the definition of disintegration (section **??**), we can write

$$\mathbb{P}[\mathsf{XDY}]^{ijk} = \mathbb{P}[\mathsf{X}]^i \mathbb{P}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{P}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \tag{76}$$

Fix some $h \in H$ and some strictly positive strategy $\mathbf{S}$ and define $\mathbf{T} : H \times D \to \Delta(X \times Y)$ by

$$\mathbf{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^l_{kj} \tag{77}$$

Note that because $\mathbf{S}$ is strictly positive and by assumption $\mathbf{S} = \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]$, $\mathbb{P}_{h,\mathbf{S}}[\mathsf{D}]$ is also strictly positive. Therefore $\mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{D}]$ is unique and therefore $\mathbf{T}$ is also unique.

Define X' and Y' by $\mathsf{X}'_{xy} = x$ and $\mathsf{Y}'_{xy} = y$. Define H' and D' by $\mathsf{H}'_{hd} = h$ and $\mathsf{D}'_{hd} = d$.

We then have

$$\begin{aligned}
\mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj} &= \mathbf{T}\underline{\mathsf{X}'}^k_{hj} && (78)\\
&= \sum_l \mathbf{T}^{kl}_{hj} && (79)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^k && (80)\\
&= \mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^k_{hj'} && (81)
\end{aligned}$$

Thus $\mathsf{X}' \perp\!\!\!\perp_{\mathbf{T}} \mathsf{D}'|\mathsf{H}'$ and so $\mathbf{T}[\mathsf{X}'|\mathsf{H}']$ exists (section 2.10) and $(\mathbf{T}, \mathsf{H}', \mathsf{D}', \mathsf{X}', \mathsf{Y}')$ is a see-do model.

Applying Equation 76 to $\mathbb{P}_{h,\mathbf{S}}$:

$$\begin{aligned}
\mathbb{P}_{h,\mathbf{S}}[\mathsf{XDY}]^{ijk} &= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]^j_i \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} && (82)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i \mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} && (83)\\
&= \mathbb{P}_{h,\mathbf{S}}[\mathsf{D}|\mathsf{X}]^j_i \mathbf{T}[\mathsf{X}'\mathsf{Y}'|\mathsf{H}'\mathsf{D}']^{ik}_{hj} && (84)\\
&= \mathbf{S}^j_i \mathbf{T}[\mathsf{X}'\mathsf{Y}'|\mathsf{H}'\mathsf{D}']^{ik}_{hj} && (85)\\
&= \mathbf{S}^j_i \mathbf{T}[\mathsf{X}'|\mathsf{H}'\mathsf{D}']^i_{hj} \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} && (86)\\
&= \mathbf{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbf{S}^j_i \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} && (87)
\end{aligned}$$

Consider some arbitrary alternative strategy $\mathbf{Q}$. By assumption

$$\mathbb{P}_{h,\mathbf{S}}[\mathsf{X}]^i = \mathbb{P}_{h,\mathbf{Q}}[\mathsf{X}]^i \tag{88}$$

$$\mathbb{P}_{h,\mathbf{S}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} = \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}]^k_{ij} \text{ for some version of } \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}] \tag{89}$$

It follows that, for some version of $\mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}]$,

$$\mathbf{T}^{kl}_{hj} = \mathbb{P}_{h,\mathbf{Q}}[\mathsf{X}]^k \mathbb{P}_{h,\mathbf{Q}}[\mathsf{Y}|\mathsf{XD}]^l_{kj} \tag{90}$$

Then by substitution of $\mathbf{Q}$ for $\mathbf{S}$ in Equation 82 and working through the same steps

$$\mathbb{P}_{h,\mathbf{S}}[\mathsf{XDY}]^{ijk} = \mathbf{T}[\mathsf{X}'|\mathsf{H}']^i_h \mathbf{Q}^j_i \mathbf{T}[\mathsf{Y}'|\mathsf{X}'\mathsf{H}'\mathsf{D}']^k_{ihj} \tag{91}$$

As $\mathbf{Q}$ was arbitrary, this holds for all strategies. □

References

G. Chiribella, Giacomo D'Ariano, and P. Perinotti. Quantum Circuit Architecture. *Physical review letters*, 101:060401, September 2008. doi: 10.1103/PhysRevLett.101.060401.

A. P. Dawid. Causal Inference without Counterfactuals. *Journal of the American Statistical Association*, 95(450):407–424, June 2000. ISSN 0162-1459. doi: 10.1080/01621459.2000.10474210. URL `https://www.tandfonline.com/doi/abs/10.1080/01621459.2000.10474210`. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.2000.10474210.

M. A. Hernán and S. L. Taubman. Does obesity shorten life? The importance of well-defined interventions to answer causal questions. *International Journal of Obesity*, 32(S3):S8–S14, August 2008. ISSN 1476-5497. doi: 10.1038/ijo.2008.82. URL `https://www.nature.com/articles/ijo200882`.

Bart Jacobs, Aleks Kissinger, and Fabio Zanasi. Causal Inference by String Diagram Surgery. In Mikoaj Bojaczyk and Alex Simpson, editors, *Foundations of Software Science and Computation Structures*, Lecture Notes in Computer Science, pages 313–329. Springer International Publishing, 2019. ISBN 978-3-030-17127-8.

Alfred Korzybski. *Science and sanity; an introduction to Non-Aristotelian systems and general semantics*. Lancaster, Pa., New York City, The International Non-Aristotelian Library Publishing Company, The Science Press Printing Company, distributors, 1933. URL `http://archive.org/details/sciencesanityint00korz`.

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

Judea Pearl. Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes. *Journal of Causal Inference*, 6(2), 2018. doi: 10.1515/jci-2018-2001. URL https://www.degruyter.com/view/j/jci.2018.6.issue-2/jci-2018-2001/jci-2018-2001.xml.

James M. Robins and Sander Greenland. Causal Inference Without Counterfactuals: Comment. *Journal of the American Statistical Association*, 95 (450):431–435, 2000. ISSN 0162-1459. doi: 10.2307/2669381. URL http://www.jstor.org/stable/2669381. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

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

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

Eyal Shahar. The association of body mass index with health outcomes: causal, inconsistent, or confounded? *American Journal of Epidemiology*, 170(8):957–958, October 2009. ISSN 1476-6256. doi: 10.1093/aje/kwp292.

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.

Appendix: